PRESERVING WHAT MATTERS: EFFICIENT-CODING-INSPIRED INFORMATION BOTTLENECK FOR WHOLE-BRAIN DECODING AND VOXEL ATTRIBUTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Vision-language pretrained models have greatly advanced brain decoding, but most decoders remain opaque, offering limited evidence about which patterns in fMRI activity influence the decoding decision. We propose an interpretable whole-brain decoding framework inspired by the efficient coding principle of *preserving what matters*. Concretely, we encourage brain-VLPM alignment through contrastive learning with variational Information Bottleneck regularization to compress redundant ROI-free whole-brain fMRI activity. Built on the whole-brain decoder, our method further explains individual decoding decisions through bottleneck-based attribution, producing voxel-level saliency maps for brain-image/text alignment. Experiments on the Natural Scenes Dataset show strong retrieval performance and quantitatively evaluated interpretability. Through ROI localization and voxel ablation analyses, the highly ranked voxels are shown to strongly influence the trained decoder's alignment scores: removing them rapidly degrades the alignment score, whereas inserting them rapidly restores and even can exceed the full-input alignment score. These results establish our framework as a practical approach to identifying informative voxels supporting individual decoding decisions.

1 INTRODUCTION

Recent advances in Vision-Language Pretrained Models (VLPMs), such as CLIP (Radford et al., 2021), have substantially reshaped the landscape of brain decoding. By aligning fMRI activity with the latent spaces of powerful multimodal models, recent studies have moved beyond early coarse-grained visual category classification toward fine-grained image retrieval (Lin et al., 2022; Gong et al., 2024b), natural image reconstruction (Takagi & Nishimoto, 2023; Ozcelik & VanRullen, 2023; Scotti et al., 2023; 2024; Li et al., 2025), caption generation (Ferrante et al., 2023; Ren et al., 2024; Xia et al., 2024), video decoding (Chen et al., 2023; Gong et al., 2024a), language decoding (Ye et al., 2025), brain agents (Qiu et al., 2025), and associated visual question answering (Shen et al., 2024). Despite these impressive achievements, most brain decoders remain opaque black boxes: they can produce plausible outputs, but rarely explain which aspects of the fMRI input support their predictions. This opacity weakens their scientific value and has led to concerns that brain-reading techniques may become "wishful thinking" without transparent decision evidence rather than reliable tools for neuroscience (Vigotsky et al., 2024).

> *Which patterns in fMRI activity influence the current decoding decision, and how much does each contribute?*

Elucidating the decision basis of brain decoders faces two critical challenges that have been largely overlooked in previous performance-driven studies. First, identifying the voxel patterns that support a decoding decision is challenging because whole-brain fMRI is high-dimensional and highly redundant. Existing methods often restrict the input to predefined Regions of Interest (ROIs) to reduce dimensionality, but this restriction may exclude distributed voxels that influence the decoding result (Epp et al., 2025). Second, existing attribution methods commonly explain label-based predictions or features within image and text inputs. Modern brain decoding instead requires attributing continuous brain–VLPM alignment directly to fMRI voxels, without discrete class labels or a fixed semantic vocabulary.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

In this work, we approach brain decoder interpretability from the perspective of efficient information flow. The efficient coding principle suggests that biological neural systems tend to represent task-relevant information with minimal redundancy (Barlow et al., 1961; Simoncelli & Olshausen, 2001; Olshausen & Field, 2004). Inspired by this principle, we incorporate the Information Bottleneck to unify representation learning and decision attribution through a shared trade-off between information compression and alignment preservation.

We first develop an ROI-free whole-brain decoder that uses bottleneck regularization to retain alignment-relevant information from high-dimensional fMRI activity. Furthermore, we apply the same bottleneck principle to formulate decision attribution as a perturbation-based optimization through the frozen decoder. Instead of treating saliency maps as by-products of gradients or attention scores, we use a learnable voxel mask to perturb the whole-brain input and regulate the information flowing through the decoder. The optimized mask produces a sample-specific saliency map highlighting voxels important for the target alignment.

We evaluate the proposed framework on the Natural Scenes Dataset (NSD). Experiments show strong cross-modal retrieval performance and voxel-level saliency maps. Beyond qualitative visualization, we evaluate the attributions at two complementary levels: macro-level agreement with the visual ROI and micro-level influence of ranked voxels on the decoder's alignment. Deleting highly ranked voxels rapidly degrades alignment, whereas inserting them restores or can exceed the full-input score. These findings suggest that our method highlights a compact voxel subset that supports alignment with the paired target.

Our contributions are summarized as follows:

- We unify whole-brain decoding and decision attribution within an Information Bottleneck framework inspired by efficient coding to preserve informative voxels while suppressing redundancy.
- We build an ROI-free whole-brain decoder with efficient-coding-inspired bottleneck regularization to learn compact, alignment-relevant decoder representations from redundant fMRI activity.
- We propose a bottleneck-based attribution method that estimates voxel-level saliency maps by preserving informative voxels for the target alignment while suppressing redundant input information.
- We introduce ROI localization and voxel ablation metrics for quantitative evaluation, demonstrating on NSD that our method achieves strong retrieval performance and identifies voxels whose insertion rapidly restores and even can exceed the full-input alignment score.

2 Efficient Coding Motivation and Information-Theoretic Formulation

The efficient coding principle posits that sensory systems represent relevant information while limiting redundancy (Barlow et al., 1961; Simoncelli & Olshausen, 2001). In our framework, this principle provides the computational motivation for controlling how much input information a decoder retains.

We operationalize this motivation with the Information Bottleneck, which provides a tractable objective for balancing compression and prediction in the learned decoder representation. Formally, given an input X and target Y, the goal is to learn a representation Z that preserves the information about Y while discarding irrelevant details from X:

$$\max \; I(Z;Y) - \beta I(Z;X), \tag{1}$$

where $I(\cdot;\cdot)$ denotes mutual information and β gives a trade-off between fitting and compression. Eq. 1 provides a computational instantiation of the compression-sufficiency trade-off motivated by efficient coding. The maximization of $I(Z;Y)$ encourages *sufficiency* for the target, while the minimization of $I(Z;X)$ encourages *efficiency* by reducing information inherited from the input.

To enable tractable computation, the objective is approximated via the Variational Information Bottleneck (VIB) (Alemi et al., 2017) by maximizing the Evidence Lower Bound (ELBO) (Kingma & Welling, 2013):

$$\max_{\theta,\phi} \; \mathbb{E}_{p(x,y)} \left[\underbrace{\mathbb{E}_{p_\theta(z|x)}[\log q_\phi(y|z)]}_{\text{sufficiency}} - \beta \underbrace{\mathbb{D}_{\mathrm{KL}}[p_\theta(z|x)||r(z)]}_{\text{efficiency}} \right], \tag{2}$$

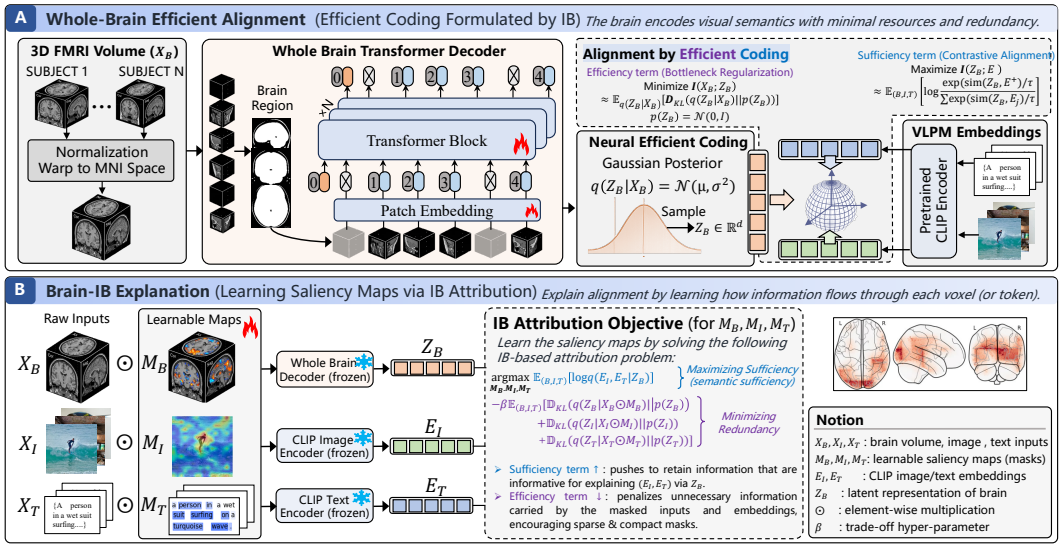


Figure 1: The overview framework diagram of the method. For alignment, the whole-brain decoder maps fMRI activity to latent representations, which are aligned with VLPM embeddings via contrastive learning while enforcing efficient coding through IB regularization. For interpretation, the IB-based decision attribution module learns a saliency map to identify the most important voxels for the decoding decision by optimizing a perturbation to maximize sufficiency while minimizing redundancy.

where $\mathbb{D}_{\mathrm{KL}}$ denotes the Kullback-Leibler divergence (Kullback & Leibler, 1951) (detailed in Appendix G.1). The target entropy $H(Y)$ is omitted because it is independent of θ and ϕ. This term acts as a tractable efficiency surrogate that reduces redundancy in the learned decoder representation.

3 Methodology

We formulate brain decoder attribution as estimating how strongly each voxel influences brain decoder's alignment with VLPM embeddings. Inspired by the efficient coding principle, we incorporate the Information Bottleneck (IB) to suppress redundant whole-brain fMRI activity while retaining alignment-relevant information. Our framework includes an ROI-free whole-brain decoder and a bottleneck-based attribution module that produces voxel-level saliency maps for individual decoding decisions.

3.1 Whole-Brain Efficient Representation

Multimodal brain decoding commonly aligns fMRI activity with the representation space of VLPMs through contrastive learning. Inspired by the efficient coding principle, we introduce a variational information bottleneck into the whole-brain decoder to focus on task-relevant information while suppressing redundant variations in high-dimensional fMRI activity.

Specifically, we model the brain decoder as a stochastic mapping $p_\theta(Z_{\mathrm{B}}|X_{\mathrm{B}})$ from fMRI volume X_{B} to a decoder representation Z_{B} in the VLPM embedding space. The training objective follows the IB formulation in Eq. 1, balancing two terms inspired by efficient coding:

- *Sufficiency term* ($I(Z;Y)$): maximizing the mutual information between decoder representations Z_{B} and VLPM embeddings Y via contrastive alignment, ensuring that these representations retain the essential semantic information useful for cross-modal alignment.
- *Efficiency term* ($-\alpha I(Z;X)$): minimizing the mutual information between Z_{B} and input brain activity X_{B} through KL divergence regularization to discourage redundant or noisy variation in whole-brain activity.

Whole-Brain Decoder. Given a 3D fMRI volume $X_{\mathrm{B}}^{s} \in \mathbb{R}^{H_s \times W_s \times D_s}$ of subject s in the native anatomical space, where H_s, W_s, and D_s denote the subject-specific height, width, and depth, accounting for variations in brain size across individuals, we first spatially normalize it to the standard *MNI152NLin6Asym* space with 2mm isotropic resolution, resulting in $X_{\mathrm{B}} \in \mathbb{R}^{H \times W \times D}$. We then partition it into non-overlapping 3D patches of size P, resulting in $L = \frac{HWD}{P^3}$ tokens. Each token is flattened and projected into a d-dimensional embedding space via a linear layer, yielding token embeddings $\{x_i\}_{i=1}^{L}$, where $x_i \in \mathbb{R}^d$. To maintain spatial structure, we add learnable positional embeddings to the sequence and prepend a learnable class token (`[CLS]`) to aggregate global information. We employ stacked Transformer layers with multi-head self-attention (Vaswani et al., 2017) to process this sequence, enabling the model to capture long-range dependencies across the whole brain. Finally, the `[CLS]` token is used to parameterize a Gaussian distribution, from which the decoder representation $Z_{\mathrm{B}} \in \mathbb{R}^d$ is sampled through reparameterization.

Contrastive Alignment (Sufficiency term). To align decoder representations with VLPM embeddings, we maximize a contrastive lower bound of the mutual information $I(Z;Y)$ (Oord et al., 2018), with the derivation provided in Appendix G.2. Given a batch of N image-text pairs and corresponding brain activities, we use the pretrained CLIP model to obtain image and text embeddings, denoted as $(Y_{\mathrm{I}}, Y_{\mathrm{T}}) \in \mathbb{R}^{N \times d}$. Let $s_{ij}^{M} = \mathrm{sim}(z_i, y_{M,j})$ denote the cosine similarity between the i-th fMRI embedding and the j-th image/text embedding. The whole-brain decoder is trained with a bidirectional contrastive objective:

$$\mathcal{L}_{\mathrm{align}} = -\frac{1}{4N} \sum_{M \in \{\mathrm{I},\mathrm{T}\}} \sum_{i=1}^{N} \left[\log \frac{\exp(s_{ii}^{M}/\tau)}{\sum_{j=1}^{N} \exp(s_{ij}^{M}/\tau)} + \log \frac{\exp(s_{ii}^{M}/\tau)}{\sum_{j=1}^{N} \exp(s_{ji}^{M}/\tau)} \right]. \tag{3}$$

Here, τ is the temperature parameter fixed to match the CLIP temperature. This objective jointly optimizes brain-image and brain-text alignment in both directions, encouraging fMRI embeddings to be close to the corresponding semantic anchors to ensure semantic sufficiency.

Information Compression (Efficiency term). Whole-brain fMRI contains variation that is not required for the current brain–VLPM alignment objective. To constrain the information that the decoder retains from these inputs, we instantiate the *efficiency term* by minimizing a variational upper bound on the mutual information $I(Z_{\mathrm{B}}; X_{\mathrm{B}})$ using the Kullback-Leibler (KL) divergence. Formally, we constrain the posterior distribution of the fmri representation $p_\theta(Z_{\mathrm{B}}|X_{\mathrm{B}})$ to match a fixed prior distribution $r(Z_{\mathrm{B}})$, typically chosen as a standard spherical Gaussian $\mathcal{N}(0, I)$. The optimization objective is defined as:

$$\mathcal{L}_{\mathrm{KL}} = \mathbb{E}_{p(x_{\mathrm{B}})}[\mathbb{D}_{\mathrm{KL}}[p_\theta(Z_{\mathrm{B}}|x_{\mathrm{B}})||r(Z_{\mathrm{B}})]] \geq I(Z_{\mathrm{B}}; X_{\mathrm{B}}). \tag{4}$$

This expected KL divergence provides a variational upper bound on representation mutual information. It penalizes the information complexity of Z_{B}, encouraging the decoder to discard input variation unnecessary for alignment and learn a compact alignment-relevant representation.

Overall Objective. The final training objective integrates the alignment and efficient coding to strictly balance semantic fidelity and representational compactness:

$$\mathcal{L}_{total} = \mathcal{L}_{align} + \alpha \mathcal{L}_{\mathrm{KL}}, \tag{5}$$

where $\alpha \geq 0$ acts as a Lagrange multiplier governing the bottleneck bandwidth. A higher α imposes a stricter information constraint, compelling the model to filter out redundancy, whereas a lower α relaxes this constraint to prioritize the maximization of semantic mutual information.

3.2 Bottleneck-based Decision Attribution

To interpret the trained brain decoder, we propose a bottleneck-based attribution mechanism that identifies informative voxels for alignment with the paired image and text embeddings. We extend unimodal (Schulz et al., 2020) and multi-modal (Wang et al., 2023) IB attribution by treating the frozen decoder as a fixed information channel. A learnable mask perturbs the fMRI input to preserve its alignment with paired image and text embeddings while reducing unnecessary information, producing voxel attribution scores and a corresponding saliency map for each decoding instance.

Problem Formulation. Formally, given a pretrained whole-brain decoder f that produces brain embeddings Z_{B} aligned with image embeddings Y_{I} and text embeddings Y_{T}, we learn a continuous

preservation mask M, parameterized by θ_M, to regulate the information retained from the fMRI input X_{B}. Rather than directly zeroing the excluded voxels, the mask defines a stochastic perturbation:

$$\widetilde{X}_{\mathrm{B}} = M \odot X_{\mathrm{B}} + (1 - M) \odot \epsilon, \qquad Z'_{\mathrm{B}} = f(\widetilde{X}_{\mathrm{B}}), \quad \epsilon \sim \mathcal{N}(0, I). \tag{6}$$

The mask therefore induces a conditional distribution $p_{\theta_M}(\widetilde{X}_{\mathrm{B}}|X_{\mathrm{B}})$ before the perturbed input is mapped to Z'_{B} by the frozen decoder. We optimize this information flow using the following bottleneck objective:

$$\max_{\theta_M} \; I(Z'_{\mathrm{B}}; Y_{\mathrm{I}}, Y_{\mathrm{T}}) \; - \; \beta \, I(\widetilde{X}_{\mathrm{B}}; X_{\mathrm{B}}), \tag{7}$$

where β controls the trade-off between preserving the current brain–image/text alignment and compressing information inherited from the fMRI input. **Maximizing Sufficiency.** The fitting term $I(Z'_{\mathrm{B}}; Y_{\mathrm{I}}, Y_{\mathrm{T}})$ measures the information that the perturbed fMRI representation retains about the paired image and text embeddings. We maximize its variational lower bound to preserve alignment with both targets:

$$I(Z'_{\mathrm{B}}; Y_{\mathrm{I}}, Y_{\mathrm{T}}) \geq \mathbb{E}_{p_{\theta_M}(z'_{\mathrm{B}}, y_{\mathrm{I}}, y_{\mathrm{T}})} \big[\log q_\phi(Y_{\mathrm{I}}, Y_{\mathrm{T}}|Z'_{\mathrm{B}}) \big] + H(Y_{\mathrm{I}}, Y_{\mathrm{T}}). \tag{8}$$

Because $H(Y_{\mathrm{I}}, Y_{\mathrm{T}})$ is independent of θ_M, it is omitted when optimizing the attribution mask. In practice, we use empirical samples of x_{B}, Y_{I}, and Y_{T} to approximate the variational optimization objective $\hat{p}(x_{\mathrm{B}}, y_{\mathrm{I}}, y_{\mathrm{T}}) \doteq \frac{1}{N} \sum_{i=1}^{N} \delta(x_{\mathrm{B}} - x_{\mathrm{B}}^i)\delta(y_{\mathrm{I}} - y_{\mathrm{I}}^i)\delta(y_{\mathrm{T}} - y_{\mathrm{T}}^i)$. As a modeling assumption, we factorize the variational distribution as $q_\phi(y_{\mathrm{I}}, y_{\mathrm{T}}|z'_{\mathrm{B}}) = q_\phi(y_{\mathrm{I}}|z'_{\mathrm{B}})q_\phi(y_{\mathrm{T}}|z'_{\mathrm{B}})$. Thus, the log of the Gaussian probability density simplifies and is proportional to the cosine similarity, giving the optimization objective of the fitting term (derived in Appendix G.3):

$$\max_{\theta_M} \mathbb{E}_{p_{\theta_M}(\widetilde{x}_{\mathrm{B}}|x_{\mathrm{B}})} \big[\cos(y_{\mathrm{I}}, f(\widetilde{x}_{\mathrm{B}})) + \cos(y_{\mathrm{T}}, f(\widetilde{x}_{\mathrm{B}})) \big]. \tag{9}$$

Minimizing Redundancy. By minimizing the efficiency term $I(\widetilde{X}_{\mathrm{B}}; X_{\mathrm{B}})$, we limit how much information the perturbed input retains from the original fMRI signal. Under Eq. 6, the conditional distribution is $p_{\theta_M}(\widetilde{X}_{\mathrm{B}}|X_{\mathrm{B}}) = \mathcal{N}(M \odot X_{\mathrm{B}}, \mathrm{diag}((1 - M)^2))$. We minimize a variational upper bound given by its KL divergence from the fixed standard Gaussian prior $r(\widetilde{X}_{\mathrm{B}}) = \mathcal{N}(0, I)$:

$$\min_{\theta_M} \mathbb{E}_{p(x_{\mathrm{B}})} \left[\mathbb{D}_{\mathrm{KL}}[p_{\theta_M}(\widetilde{x}_{\mathrm{B}}|x_{\mathrm{B}})||r(\widetilde{x}_{\mathrm{B}})] \right]. \tag{10}$$

Final Tractable Objective. The preceding bounds define a variational surrogate over the data distribution. For attribution, we optimize its sample-wise form separately for each fMRI–image–text triplet, keeping the decoder and paired targets fixed:

$$\theta_M^* = \arg\max_{\theta_M} \; I_{\mathrm{lower}}(Z'_{\mathrm{B}}; Y_{\mathrm{I}}, Y_{\mathrm{T}}) - \beta I_{\mathrm{upper}}(\widetilde{X}_{\mathrm{B}}; X_{\mathrm{B}}). \tag{11}$$

The perturbation expectation is approximated using independently sampled input noise. Optimizing this objective retains voxel information that preserves the decoder's alignment while suppressing information with less influence on that alignment. The optimized voxel-wise KL values define the voxel attribution scores, whose spatial arrangement constitutes the saliency map for the current decoding instance.

4 Empirical Evaluation

4.1 Dataset.

We evaluate on the *Natural Scenes Dataset* (NSD) (Allen et al., 2022), where functional magnetic resonance imaging (fMRI) was collected from 8 subjects. Each subject viewed 9,000 unique images (treated as $3 \times 9,000$ independent samples for training) and 1,000 shared images (averaged across 3 repetitions for testing) over 40 sessions.

Brain. We normalize the trial-wise betas (estimated via GLMdenoise with voxel-wise hemodynamic response function (HRF) and ridge regression) from the native functional space (`func1mm`) to the `MNI152NLin6Asym 2mm` space, resulting in a volume dimension of $(91, 109, 91)$ as model input. We perform voxel-wise z-score normalization to standardize the fMRI volumes and apply the `MNI152_T1_2mm` brain mask to exclude non-brain patches.

Image. We use the original images from the NSD, which are resized to 224×224 pixels and normalized.

Text. For model training, we used 4-bit quantized Qwen2.5-VL-32B to generate detailed descriptions. These descriptions provide richer semantic information and facilitate brain-text alignment. We separately ablate COCO captions and Qwen-generated descriptions in Table 2, and further evaluate their retrieval performance and semantic consistency in Appendix H. COCO captions serve as the default text targets for attribution.

4.2 Evaluation Metrics

Alignment Evaluation. We evaluate cross-modal alignment via bidirectional retrieval tasks on the NSD test set. Specifically, we compute cosine similarities between fMRI and VLPM embeddings. Following the protocol in (Scotti et al., 2023), we sample a random subset of 300 pairs from the full 1,000-sample test set and repeat this process 30 times to ensure statistical robustness. We report the average Top-K Recall (R@K) for both brain-to-modality (Image/Text) and modality-to-brain retrieval (chance=K/300).

Explanation Evaluation. We introduce two quantitative metrics to evaluate the agreement of saliency maps with the visual ROI and their ability to identify voxels that affect decoder alignment.

a. Macro-level Localization: We evaluate how well the voxel attribution scores distinguish visually responsive regions from the rest of the brain, using the binary nsdgeneral ROI mask as the reference. At each threshold T, we binarize the saliency map to obtain M_T and compute the true positive rate (TPR) and false positive rate (FPR) relative to this mask. Varying T traces the receiver operating characteristic (ROC) curve, whose area under the curve (AUC) summarizes localization performance across thresholds:

$$\mathrm{TPR}(T) = \frac{\sum_{v\in\mathcal{V}} \mathbb{1}_{M_T^v \wedge S_{\mathrm{ROI}}^v}}{\sum_{v\in\mathcal{V}} \mathbb{1}_{S_{\mathrm{ROI}}^v}}, \quad \mathrm{FPR}(T) = \frac{\sum_{v\in\mathcal{V}} \mathbb{1}_{M_T^v \wedge (S_{\mathrm{ROI}}^-)^v}}{\sum_{v\in\mathcal{V}} \mathbb{1}_{(S_{\mathrm{ROI}}^-)^v}}, \tag{12}$$

where $\mathcal{V}$ indexes valid brain voxels, S_{ROI} is the binary mask of the nsdgeneral ROI and S_{ROI}^- is its complement within $\mathcal{V}$. The $\mathbb{1}$ is the indicator function.

b. Micro-level Ablation: Fine-grained saliency evaluation is challenging because voxel-wise explanation labels are unavailable. We therefore introduce the ablation-based metric (Petsiuk et al., 2018) that measures how the brain decoder's alignment score changes when voxels are removed or restored in descending order of voxel attribution scores. For deletion, we progressively replace top-ranked voxels with the zero baseline, corresponding to the training-set mean after normalization. For insertion, we start from the same zero baseline volume and progressively restore top-ranked voxels from the original input. We report the cosine similarity ratio after/before ablation and summarize the curve by AUC. A more informative ranking by voxel attribution scores is expected to produce a lower deletion AUC and a higher insertion AUC.

4.3 Empirical Results

Quantitative Evaluation of Alignment. We compare the bidirectional retrieval of our brain decoder with several methods, including Ridge regression (Takagi & Nishimoto, 2023), Mindeye (Scotti et al., 2023), Mindeye2 (Scotti et al., 2024), and TGBD (Kong et al., 2025). We adapt their brain encoders to the same image and text retrieval objective under the controlled configuration detailed in Appendix C. As shown in Fig. 2, our method performs strongly in both brain-image and brain-text retrieval across all subjects, demonstrating the effectiveness of incorporating the efficient coding principle via the information bottleneck. Notably, our method achieves substantial improvements in R@1, even approaching the R@5 of other methods. Our R@5 and R@10 substantially outperform the others. This indicates that the learned brain embeddings capture rich semantic information aligned with VLPMs. This lays a solid foundation for our subsequent attribution analysis.

Quantitative Evaluation of Decision Attribution. We compare our methods with Saliency Map (Simonyan et al., 2013), RISE (Petsiuk et al., 2018), GradCAM (Selvaraju et al., 2017), the method of Chefer et al. (2021), and Mfaba (Zhu et al., 2024). We assess two complementary properties: agreement with the visual ROI and the influence of ranked voxels on the decoder's alignment score.

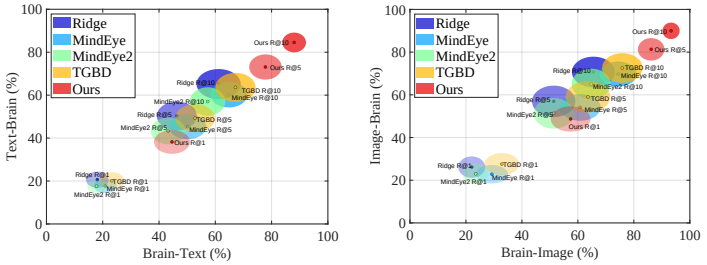


Figure 2: Brain-VLPM alignment performance is measured by bidirectional retrieval accuracy. The horizontal axis reports forward retrieval (retrieval of 'modality' via brain activity), while the vertical axis reports backward (retrieval of brain activity by 'modality'). We show the bidirectional retrieval accuracy of brain-text (left) and brain-image (right). The center point is the average of the 8 subjects, while the patch represents the overall distribution of the 8 subjects. Performance is better towards the top-right corner.

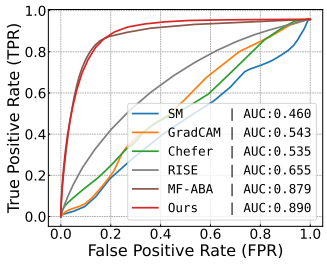


Figure 3: Macro-level ROI localization evaluated using ROC curves. Curves closer to the upper-left corner indicate better discrimination between voxels inside and outside the nsdgeneral visual ROI based on their voxel attribution scores.

Macro-level ROI localization. Fig. 3 evaluates how well the saliency maps localize the nsdgeneral visual ROI. Our ROC curve lies close to the upper-left corner and achieves the largest area among the compared methods, indicating that the saliency maps effectively distinguish voxels within this ROI from those outside it. This localization metric establishes agreement with a broad visually responsive region. Determining whether the voxel rankings capture information specific to an individual sample requires a finer-grained evaluation.

Micro-level voxel deletion and insertion. To evaluate whether the saliency maps rank voxels according to their influence on the decoder's paired fMRI-image/text alignment, beyond a shared preference for visual regions, we introduce two baselines in Fig. 4. *Random* deletes or inserts valid voxels with a random order, providing a baseline without saliency guidance. *Shuffle* deletes or inserts valid voxels using the saliency ranking from another sample. This breaks the correspondence between the ranking and the fMRI sample while preserving broad visual-region preference.

Random produces gradual, approximately linear changes in alignment, consistent with informative voxels being distributed throughout the perturbation sequence. Shuffle yields faster degradation and recovery than Random, consistent with its broad visual-region preference. Our saliency maps produce substantially steeper initial changes and even exhibit an overshoot, with alignment scores exceeding those obtained using the full input. These patterns suggest that our method identifies a compact, informative voxel subset supporting the target alignment. This subset can yield stronger target alignment than the complete input, suggesting that subsequently restored voxels may dilute the alignment. Together with the controls, these effects support the selective preservation of sample-specific alignment information.

Fig. 5 compares existing attribution methods under the same evaluation protocol. Our method induces faster degradation and recovery, achieving the lowest deletion AUC and highest insertion AUC for both image and text alignment. These results demonstrate that our method more effectively identifies voxels that strongly influence the target alignment.

4.4 Ablation Study and Parameter Sensitivity

Our framework uses the Information Bottleneck to unify brain decoding and decision attribution. In the decoder, IB provides a computational model of efficient coding that preserves task-relevant information from whole-brain activity. During attribution, IB is applied to the frozen decoder to identify voxels with a strong influence on the current brain-VLPM alignment. We examine its role and sensitivity to compression strength at both stages.

4.4.1 Information Bottleneck Decoder

Table 1 shows that removing IB reduces bidirectional retrieval accuracy. Replacing the KL term with L1 regularization to encourage sparse brain embeddings offers only limited improvement over the

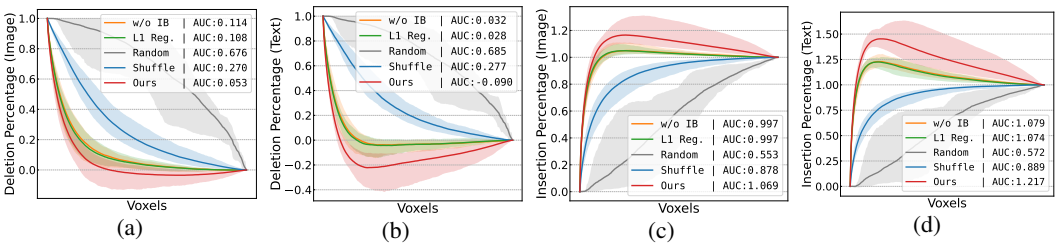


Figure 4: Voxel deletion/insertion with random rankings (Random), voxel rankings transferred from other samples (Shuffle), and our attribution procedure applied to decoders trained without IB, with L1 regularization, or with IB (Ours). Panels (a,b) show deletion and (c,d) show insertion for image and text alignment.

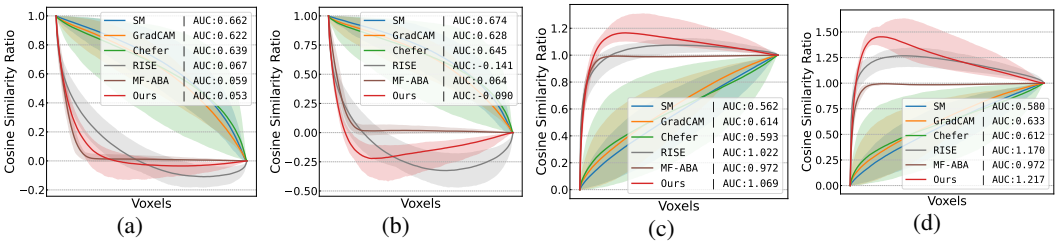


Figure 5: Comparison different attribution methods for micro-level ablation metrics. The horizontal axis denotes the voxels deleted (inserted) based on the saliency map, while the vertical axis reports the cosine similarity ratio of after/before deletion (insertion). (a-d) deletion/insertion voxels for brain-image alignment (a,c) and brain-text alignment (b,d).

decoder without IB. The IB decoder performs best in all four directions, establishing an empirical benefit for alignment beyond what L1 sparsity regularization provides. We interpret this benefit through information compression, consistent with the efficient coding motivation. The mutual information estimates provide further insight, remaining high without IB and with L1 regularization but decreasing substantially with IB. Better retrieval with less retained input information suggests that IB suppresses redundancy while preserving decoding-relevant information.

This compression must remain compatible with semantic sufficiency. Increasing α progressively reduces estimated mutual information, whereas retrieval peaks at $\alpha = 10^{-5}$ and declines under a stronger constraint. The resulting trade-off indicates that moderate compression benefits alignment, but excessive compression discards task-relevant information.

4.4.2 Information Bottleneck-based Decision Attribution

The contribution of decoder-side IB extends to decision attribution. With the attribution procedure held fixed, the IB decoder exhibits faster alignment degradation and recovery than the decoders trained without IB or with L1 regularization (Fig. 4). These results suggest that IB training concentrates the decoder's reliance on a compact voxel subset identified by the attribution procedure.

Within the attribution procedure, β controls the balance between preserving alignment and restricting information flow. Small values produce diffuse saliency maps, whereas large values suppress potentially informative voxels (Fig. 6). An intermediate value of $\beta = 0.01$ achieves the best balance, as supported by ROI localization and deletion/insertion results in Figs. 7 and 8. Additional sensitivity analyses for noise variance, learning rate, and optimization steps are provided in Appendix F.4.

5 Conclusion

In this work, we presented an interpretable whole-brain decoding framework guided by the principle of efficient coding. By introducing Information Bottleneck regularization into brain-VLPM alignment, our method learns compact fMRI representations that retain decoding-relevant information while suppressing redundant whole-brain fMRI activity. Building on this efficient-coding formulation, we

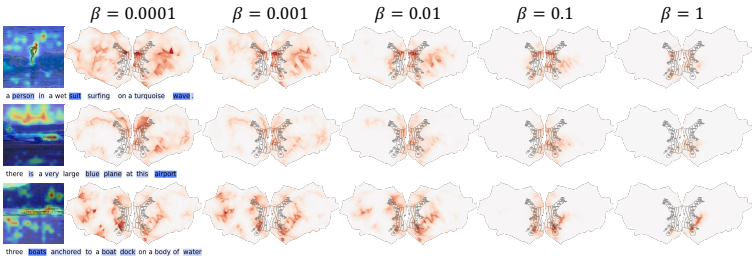


Figure 6: Examples of saliency maps of S01 obtained by different β.

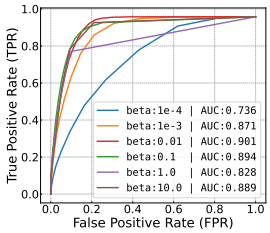
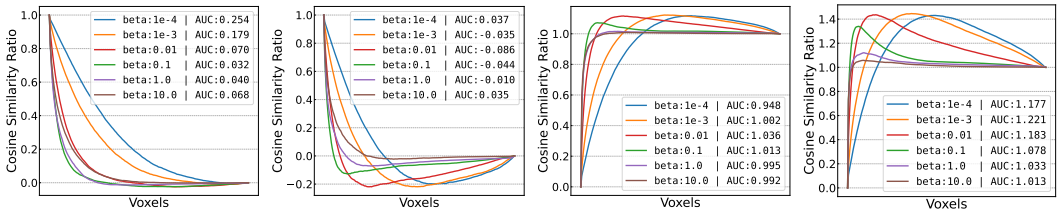


Figure 7: The ROC-AUC of macro-level localization for different β.



(a) (b) (c) (d)

Figure 8: Comparison micro-level ablation metric of different β in our method. Reporting the alignment ratio after versus before deletion or insertion (insertion). (a-d) deletion/insertion voxels for brain-image alignment (a,c) and brain-text alignment (b,d).

Table 1: Effect of Information Bottleneck on representation mutual information and bidirectional retrieval. Results are reported as mean±std across eight subjects. **Best** retrieval results are highlighted.

Methods	Mutual Information	Brain-Image	Image-Brain	Brain-Text	Text-Brain
w/o IB	745.31 ±178.90	56.24 ±2.40	45.30 ±2.35	42.66 ±2.17	36.26 ±2.24
L1 Reg.	728.21 ±143.56	56.72 ±2.47	45.42 ±2.33	42.79 ±2.40	35.65 ±2.36
$\alpha = 1e-6$	500.71 ±87.34	56.95 ±2.24	45.79 ±2.46	43.98 ±2.25	36.45 ±2.16
$\alpha = 1e-5$	57.56 ±6.87	**57.40** ±2.22	**48.68** ±2.42	**44.60** ±2.31	**38.24** ±2.22
$\alpha = 1e-4$	6.02 ±1.04	45.77 ±2.35	44.69 ±2.33	37.11 ±2.22	35.89 ±2.30

* Mutual information is estimated using the variational upper bound.

further developed a bottleneck-based attribution method that produces voxel-level saliency maps for individual decoding decisions. Experiments on the NSD demonstrate that our framework achieves strong cross-modal retrieval performance while providing quantitatively evaluated saliency maps. In particular, ROI localization characterizes agreement with a broad visual region, while voxel ablation shows that the top-ranked voxels exert a strong influence on the frozen decoder's alignment output. Removing these voxels sharply degrades alignment, whereas inserting them rapidly restores and can even surpass the full-input alignment. These results suggest that efficient coding offers a principled route toward transparent brain decoders, enabling modern brain decoding models to move beyond output generation toward decision-specific interpretation.

Limitation Our framework has several limitations. First, the reliability of the voxel attribution scores depends on the accuracy of the decoder as well as the attribution method. An inaccurate decoder may therefore produce misleading saliency maps. The saliency maps characterize the trained decoder's dependence on its input voxels and do not establish causal roles for cortical regions or a biological implementation of the Information Bottleneck. Improving decoder performance is therefore an ongoing topic. Second, while our whole-brain decoder normalized anatomical differences, inter-subject functional variability may reduce both decoding accuracy and attribution confidence. Representing and explaining shared and subject-specific representations across individuals remains an important direction for future work. Finally, extending the framework to dynamic neural data, such as EEG or MEG, could provide further insight into the temporal dynamics of brain decoding.

AI USE STATEMENT

Generative AI tools were used to assist with translation, manuscript drafting and editing, experimental design, and preparation of figures and tables. In addition, Qwen2.5-VL-32B was used to generate detailed image descriptions as auxiliary text supervision for brain–VLPM alignment; the prompting procedure, generated examples, and corresponding evaluations are documented in the appendix. All AI-assisted manuscript content was reviewed and revised by the authors. Mathematical derivations and implementations were independently checked and verified by the authors, who take full responsibility for the final content of this work.

ETHICS STATEMENT

This work performs secondary analysis of the existing Natural Scenes Dataset (NSD) and involves no new human-subject recruitment, intervention, or neuroimaging data collection. Ethical approval and informed-consent procedures for the original data collection are described in the corresponding NSD publication. The present study focuses on methodological development for brain decoding and model interpretation. The resulting voxel attribution maps characterize the dependence of the trained decoder on its fMRI inputs and should not be interpreted as evidence of causal neural mechanisms or direct clinical significance.

REPRODUCIBILITY STATEMENT

We provide detailed materials to support reproducibility throughout the main text and appendix. The main text specifies the Information Bottleneck formulation, whole-brain decoder, attribution objective, preprocessing pipeline, evaluation metrics, and experimental protocols. The appendix further provides the NSD acquisition and evaluation protocols, complete implementation and baseline settings, robustness analyses of the micro-ablation metrics, computational-cost analysis, additional subject-wise and sensitivity analyses, theoretical derivations of the information-theoretic objectives, and the complete procedure for generating and evaluating Qwen2.5-VL-32B text descriptions. All retrieval baselines are evaluated under matched data splits, preprocessing, semantic targets, and optimization settings. Anonymous source code is available at https://anonymous.4open.science/r/BrainIBA-0F9F.

REFERENCES

Alexander A Alemi, Ian Fischer, Joshua V Dillon, and Kevin Murphy. Deep variational information bottleneck. In *International Conference on Learning Representations*, 2017.

Emily J Allen, Ghislain St-Yves, Yihan Wu, Jesse L Breedlove, Jacob S Prince, Logan T Dowdle, Matthias Nau, Brad Caron, Franco Pestilli, Ian Charest, et al. A massive 7t fmri dataset to bridge cognitive neuroscience and artificial intelligence. *Nature neuroscience*, 25(1):116–126, 2022.

Horace B Barlow et al. Possible principles underlying the transformation of sensory messages. *Sensory communication*, 1(01):217–233, 1961.

Sara Cammarota, Matteo Ferrante, and Nicola Toschi. Bridging brains and concepts: Interpretable visual decoding from fmri with semantic bottlenecks. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

Hila Chefer, Shir Gur, and Lior Wolf. Generic attention-model explainability for interpreting bi-modal and encoder-decoder transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 397–406, 2021.

Zijiao Chen, Jiaxin Qing, and Juan Helen Zhou. Cinematic mindscapes: High-quality video reconstruction from brain activity. *Advances in Neural Information Processing Systems*, 36:24841–24858, 2023.

Samira M Epp, Gabriel Castrillón, Beijia Yuan, Jessica Andrews-Hanna, Christine Preibisch, and Valentin Riedl. Bold signal changes can oppose oxygen metabolism across the human cortex. *Nature neuroscience*, pp. 1–12, 2025.

Russell Epstein and Nancy Kanwisher. A cortical representation of the local visual environment. *Nature*, 392(6676):598–601, 1998.

Lang Feng, Weihao Tan, Zhiyi Lyu, Longtao Zheng, Haiyang Xu, Ming Yan, Fei Huang, and Bo An. Towards efficient online tuning of VLM agents via counterfactual soft reinforcement learning. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 16884–16903. PMLR, 2025.

Matteo Ferrante, Furkan Ozcelik, Tommaso Boccato, Rufin VanRullen, and Nicola Toschi. Brain captioning: Decoding human brain activity into images and text. *arXiv preprint arXiv:2305.11560*, 2023.

Isabel Gauthier, Pawel Skudlarski, John C Gore, and Adam W Anderson. Expertise for cars and birds recruits brain areas involved in face recognition. *Nature neuroscience*, 3(2):191–197, 2000.

Zixuan Gong, Guangyin Bao, Qi Zhang, Zhongwei Wan, Duoqian Miao, Shoujin Wang, Lei Zhu, Changwei Wang, Rongtao Xu, Liang Hu, et al. Neuroclips: Towards high-fidelity and smooth fmri-to-video reconstruction. *Advances in Neural Information Processing Systems*, 37:51655–51683, 2024a.

Zixuan Gong, Qi Zhang, Guangyin Bao, Lei Zhu, Yu Zhang, Ke Liu, Liang Hu, and Duoqian Miao. Lite-mind: Towards efficient and robust brain representation learning. In *ACM Multimedia 2024*, 2024b.

Nancy Kanwisher, Josh McDermott, and Marvin M Chun. The fusiform face area: a module in human extrastriate cortex specialized for face perception. *Journal of neuroscience*, 17(11):4302–4311, 1997.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.

Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

Xiangtao Kong, Kexin Huang, Ping Li, and Lei Zhang. Toward generalizing visual brain decoding to unseen subjects. In *The Thirteenth International Conference on Learning Representations*, 2025.

Solomon Kullback and Richard A Leibler. On information and sufficiency. *The annals of mathematical statistics*, 22(1):79–86, 1951.

Haoyu Li, Hao Wu, and Badong Chen. Neuraldiffuser: Neuroscience-inspired diffusion guidance for fmri visual reconstruction. *IEEE Transactions on Image Processing*, 34:552–565, 2025.

Sikun Lin, Thomas Sprague, and Ambuj K Singh. Mind reader: Reconstructing complex images from brain activities. *Advances in Neural Information Processing Systems*, 35:29624–29636, 2022.

Yongqiang Ma, Yulong Liu, Liangjun Chen, Guibo Zhu, Badong Chen, and Nanning Zheng. Brainclip: Brain representation via clip for generic natural visual stimulus decoding. *IEEE Transactions on Medical Imaging*, 44(10):3962–3972, 2025.

Bruno A Olshausen and David J Field. Sparse coding of sensory inputs. *Current opinion in neurobiology*, 14(4):481–487, 2004.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

Furkan Ozcelik and Rufin VanRullen. Natural scene reconstruction from fmri signals using generative latent diffusion. *Scientific Reports*, 13(1):15666, 2023.

Ankit Pensia, Varun Jog, and Po-Ling Loh. Extracting robust and accurate features via a robust information bottleneck. *IEEE Journal on Selected Areas in Information Theory*, 1(1):131–144, 2020.

Vitali Petsiuk, Abir Das, and Kate Saenko. Rise: Randomized input sampling for explanation of black-box models. In *Proceedings of the British Machine Vision Conference (BMVC)*, 2018.

Weikang Qiu, Zheng Huang, Haoyu Hu, Aosong Feng, Yujun Yan, and Rex Ying. Mindllm: A subject-agnostic and versatile model for fmri-to-text decoding. *arXiv preprint arXiv:2502.15786*, 2025.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PmLR, 2021.

Ziqi Ren, Jie Li, Xuetong Xue, Xin Li, Fan Yang, Zhicheng Jiao, and Xinbo Gao. Mindsemantix: Deciphering brain visual experiences with a brain-language model. *arXiv preprint arXiv:2405.18812*, 2024.

Karl Schulz, Leon Sixt, Federico Tombari, and Tim Landgraf. Restricting the flow: Information bottlenecks for attribution. In *International Conference on Learning Representations*, 2020.

Paul Scotti, Atmadeep Banerjee, Jimmie Goode, Stepan Shabalin, Alex Nguyen, Aidan Dempster, Nathalie Verlinde, Elad Yundler, David Weisberg, Kenneth Norman, et al. Reconstructing the mind's eye: fmri-to-image with contrastive learning and diffusion priors. *Advances in Neural Information Processing Systems*, 36:24705–24728, 2023.

Paul S Scotti, Mihir Tripathy, Cesar Kadir Torrico Villanueva, Reese Kneeland, Tong Chen, Ashutosh Narang, Charan Santhirasegaran, Jonathan Xu, Thomas Naselaris, Kenneth A Norman, et al. Mindeye2: Shared-subject models enable fmri-to-image with 1 hour of data. *arXiv preprint arXiv:2403.11207*, 2024.

Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pp. 618–626, 2017.

Guobin Shen, Dongcheng Zhao, Xiang He, Linghao Feng, Yiting Dong, Jihang Wang, Qian Zhang, and Yi Zeng. Neuro-vision to language: Enhancing brain recording-based visual reconstruction and language interaction. *Advances in Neural Information Processing Systems*, 37:98083–98110, 2024.

Eero P Simoncelli and Bruno A Olshausen. Natural image statistics and neural representation. *Annual review of neuroscience*, 24(1):1193–1216, 2001.

Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*, 2013.

Ayush Srivastava, Oshin Dutta, Jigyasa Gupta, Sumeet Agarwal, and Prathosh AP. A variational information bottleneck based method to compress sequential networks for human action recognition. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pp. 2745–2754, 2021.

Yu Takagi and Shinji Nishimoto. High-resolution image reconstruction with latent diffusion models from human brain activity. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14453–14463, 2023.

Naftali Tishby, Fernando C Pereira, and William Bialek. The information bottleneck method. *arXiv preprint physics/0004057*, 2000.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Andrew D Vigotsky, Gian Domenico Iannetti, and A Vania Apkarian. Mental state decoders: game-changers or wishful thinking? *Trends in Cognitive Sciences*, 28(10):884–895, 2024.

Ying Wang, Tim GJ Rudner, and Andrew G Wilson. Visual explanations of image-text representations via multi-modal information bottleneck attribution. *Advances in Neural Information Processing Systems*, 36:16009–16027, 2023.

Weihao Xia, Raoul de Charette, Cengiz Oztireli, and Jing-Hao Xue. Umbrae: Unified multimodal brain decoding. In *European Conference on Computer Vision*, pp. 242–259. Springer, 2024.

Ziyi Ye, Qingyao Ai, Yiqun Liu, Maarten de Rijke, Min Zhang, Christina Lioma, and Tuukka Ruotsalo. Generative language reconstruction from brain recordings. *Communications Biology*, 8 (1):346, 2025.

Zhiyu Zhu, Huaming Chen, Jiayu Zhang, Xinyi Wang, Zhibo Jin, Minhui Xue, Dongxiao Zhu, and Kim-Kwang Raymond Choo. Mfaba: A more faithful and accelerated boundary-based attribution method for deep neural networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 17228–17236, 2024.

Appendix

A Related Works

A.1 Brain-VLPM Alignment

Vision-language pretrained models have enabled brain decoding beyond category prediction, supporting image and text retrieval (Ma et al., 2025), image reconstruction (Takagi & Nishimoto, 2023; Ozcelik & VanRullen, 2023; Scotti et al., 2023; 2024), caption generation (Ferrante et al., 2023; Ren et al., 2024; Xia et al., 2024), video reconstruction (Chen et al., 2023; Gong et al., 2024a), text generation (Ye et al., 2025), and brain agents (Qiu et al., 2025). However, existing methods are often criticized as opaque black boxes (Vigotsky et al., 2024) and commonly rely on predefined visual ROIs to reduce the dimensionality of fMRI signals. We instead introduce the Information Bottleneck into whole-brain decoding, encouraging the decoder to focus on alignment-relevant voxel patterns while suppressing redundant fMRI activity.

A.2 Decision Attribution

Decision attribution identifies the input evidence supporting a model output. Gradient-based (Simonyan et al., 2013; Selvaraju et al., 2017), perturbation-based (Petsiuk et al., 2018), attention-based (Chefer et al., 2021), and boundary-based (Zhu et al., 2024) attribution methods estimate the contribution of model features, but do not explicitly identify a compact feature subset that preserves multimodal alignment. M2IB addresses this problem for VLPMs by attributing image-text correspondence to image regions and text tokens (Wang et al., 2023). We extend bottleneck-based attribution to whole-brain fMRI inputs, producing voxel attribution scores for continuous brain-image/text alignment. Recent brain interpretation methods instead explain reconstruction through predefined semantic concepts (Cammarota et al.) or cross-attention between cortical parcels and diffusion features (Feng et al., 2025). Our method directly assigns voxel attribution scores according to each voxel's influence on decoder's alignment without a fixed concept vocabulary or generative reconstruction process.

A.3 Information Bottleneck for Representation and Attribution

The Information Bottleneck formalizes representation learning as a trade-off between retaining information relevant to a target and compressing information inherited from the input (Tishby et al., 2000). Variational formulations make this objective tractable in deep networks and enable compact representations under task-dependent information constraints (Srivastava et al., 2021; Pensia et al., 2020). Information Bottleneck Attribution applies the same principle to identify a restricted set of features that remains sufficient for a model output (Schulz et al., 2020; Wang et al., 2023). We integrate these two roles of the Information Bottleneck within a unified computational framework for whole-brain multimodal decoding and decision attribution.

B NATURAL SCENES DATASET (NSD) PROTOCOLS.

The NSD comprises high-resolution whole-brain fMRI measurements from 8 healthy adult subjects (2 males, 6 females) who each participated in 30–40 scan sessions over the course of a year. Data were acquired on a 7T Siemens Magnetom scanner equipped with a 32-channel head coil, utilizing a gradient-echo EPI sequence with a high spatial resolution of 1.8mm isotropic and a repetition time (TR) of 1.6s. Each scanning session consisted of 12 runs, with each run lasting 5 minutes and containing 75 trials (62 or 63 stimulus trials with occasional interspersed blank trials). The visual stimuli, sourced from the COCO dataset (425×425 pixels), were projected at a visual angle of $8.4° \times 8.4°$. A rapid event-related paradigm was employed where each 4s trial consisted of an image presentation for 3s followed by a 1s inter-stimulus interval. To maintain vigilance, subjects performed a continuous recognition task throughout the sessions. Each unique image was presented exactly three times.

C IMPLEMENTATION DETAILS

• *Whole-Brain Decoder Model Architecture.* We instantiate the whole-brain decoder using a 3D Vision Transformer (ViT) architecture for each subject. Specifically, we replace the standard 2D patch embedding layer with a 3D patch embedding module, utilizing a patch size of $14 \times 14 \times 14$. The decoder depth is configured to 12 blocks, while other architectural hyperparameters (e.g., hidden dimension, attention heads) remain consistent with the `CLIP-ViT-H/14` configuration. The pre-trained `CLIP-ViT-H/14` is employed as the frozen VLPM teacher to provide semantic supervision.

• *Training Protocol of Whole-Brain Decoder.* The whole-brain decoder is trained on a single NVIDIA RTX A6000 (48GB) GPU for 150 epochs. We utilize the AdamW optimizer with a base learning rate of 3×10^{-4}. We employ a batch size of 256 and, to expand the negative sample pool for contrastive learning, we implement a memory bank that caches the preceding 4,096 samples. We use a OneCycle learning-rate schedule with a two-epoch warm-up, a maximum learning rate of 3×10^{-4}, and decay to 1×10^{-5}.

• *Retrieval Baseline Settings.* We reproduce Ridge, MindEye, MindEye2, and TGBD using the same NSD samples and train/test splits, voxel-wise standardized fMRI inputs in MNI152 space at 2-mm resolution, and frozen CLIP ViT-H/14 embeddings. Brain–text alignment uses the same Qwen2.5-VL-32B descriptions across methods. For task adaptation, we retain the brain-encoding backbone of each method: a linear mapping for Ridge, the four-block residual MLP of MindEye, the four-block MLP-Mixer of MindEye2, and the 3D CNN followed by a residual MLP for TGBD. Ridge, MindEye, and MindEye2 receive the vectorized valid voxels from the same whole-brain volumes, while TGBD retains their three-dimensional arrangement. We replace method-specific reconstruction or generation heads with a linear output layer matching the CLIP embedding dimension. Instead, every encoder is optimized with the same equally weighted bidirectional brain-image and brain-text contrastive losses and the same 4,096-sample memory bank.

All trainable retrieval models use AdamW for 150 epochs, a batch size of 256, and the same OneCycle schedule with a two-epoch warm-up, a maximum learning rate of 3×10^{-4}, and decay to 1×10^{-5}. No data augmentation is applied. Architecture-specific widths, depths, and dropout rates follow the corresponding implementations, while the optimization settings are fixed across methods rather than tuned separately. The frozen CLIP encoders and Qwen descriptions are shared across methods. For each of the 30 evaluation repetitions, all methods use the same 300 candidate pairs sampled from the same 1,000-sample test set, and we report the mean and standard deviation of R@K.

• *L1 Baseline Settings.* The L1 control in Table 1 retains the whole-brain decoder architecture, contrastive objectives, optimizer, and training schedule of the IB decoder. We replace the variational KL penalty with the mean absolute value of the elements in the brain embedding before L2 normalization. This regularizer encourages sparsity in the learned brain embeddings rather than directly selecting input voxels. Its coefficient is set to 10^{-5}, matching the selected IB coefficient. The comparison therefore tests a matched embedding-level L1 replacement of the IB term.

• *Saliency Map Optimization.* For each brain–image–text triplet, we freeze the trained decoder and insert the parameterized bottleneck before its embedding layer. The mask logits are defined on a $7 \times 8 \times 7$ spatial grid and initialized to 5.0. A sigmoid transformation and trilinear interpolation map

them to the resolution of the padded fMRI input. We replicate each input 10 times and independently sample standard Gaussian perturbations for the copies. The mask is optimized for 10 steps using Adam with a learning rate of 1.0 and $\beta = 0.01$. After optimization, the voxel-wise KL values are averaged across the 10 perturbed copies to obtain the voxel attribution scores. We multiply the resulting volume by the valid-brain mask, remove the spatial padding, and normalize it by its maximum absolute value. The normalized voxel attribution scores constitute the final saliency map.

D Reliable Analysis of Micro-Ablation Metrics

Deletion and insertion evaluate voxel rankings by measuring changes in target alignment through the frozen decoder. To clarify the interpretation and reliability of these metrics, we examine their normalization, sensitivity to the replacement baseline, and transfer across attribution and evaluation targets.

Metric definition and interpretation. For an input x_{B} and its paired image or text embedding y_M, let $x_{\mathrm{B}}^{(t)}$ denote the input after deleting or inserting a fraction t of valid brain voxels in descending order of voxel attribution scores. With the frozen decoder $\mathcal{E}$, the alignment ratio and its area under the curve are

$$r_M(t) = \frac{\cos\left(\mathcal{E}(x_{\mathrm{B}}^{(t)}), y_M\right)}{\cos(\mathcal{E}(x_{\mathrm{B}}), y_M)}, \qquad \mathrm{AUC}_M = \int_0^1 r_M(t)\,\mathrm{d}t, \qquad M \in \{\mathrm{I}, \mathrm{T}\}. \tag{13}$$

The reported AUC numerically integrates this ratio curve and differs from ROC-AUC. Because normalization uses the original paired similarity, neither the ratio nor its integral is bounded to $[0, 1]$. An insertion ratio above one means that the restored subset yields stronger target alignment than the full input. A negative deletion ratio, when the original similarity is positive, means that removing salient voxels reverses the sign of alignment. The overshoot is consistent with a compact voxel subset whose effect on brain decoder is diluted by subsequently restored voxels. Its interpretation depends on the controls, however, rather than on an out-of-range value alone. In Fig. 4, the stronger effects of correctly paired maps relative to Random and Shuffle show that sample-specific voxel rankings have a stronger influence beyond a shared preference for visual regions.

Robustness to the replacement baseline. Zero corresponds to the voxel-wise mean after standardization, but replacing many voxels with zero may still produce inputs outside the training distribution. We therefore repeat deletion and insertion using a mismatched test fMRI volume as the replacement baseline. Deletion substitutes the donor volume's values at the selected voxel locations, whereas insertion starts from the donor volume and progressively restores the current sample's original values. The decoder, voxel rankings, and paired targets remain unchanged. This manipulation changes the replacement values, whereas Shuffle changes the correspondence between the saliency map and the evaluated sample.

Fig. 9 preserves the pattern observed with zero replacement. Correctly paired maps induce substantially faster alignment degradation and recovery than Random and Shuffle, and their curves remain close to those obtained with zero replacement, including the insertion overshoot. These results show that the stronger influence of correctly paired voxel rankings is robust to this alternative baseline, reducing concern that they arise from zero replacement alone.

Cross-modal evaluation. Optimizing and evaluating attribution against the same alignment target can favor maps tailored to that objective. To test whether the selected voxels retain information beyond the optimization target, we optimize a saliency map using only image or only text alignment and evaluate deletion and insertion against the other modality. The trained decoder remains frozen throughout, and the evaluation modality is excluded from saliency optimization. Because concise COCO captions may omit visual details, this supplementary experiment also uses richer Qwen descriptions as text-only attribution targets for image-alignment evaluation.

As shown in Fig. 10, image-based attribution produces rapid degradation and recovery of text alignment, outperforming Random and Shuffle despite not using the text target during attribution. Transfer from text attribution to image evaluation depends on the text anchor. Maps optimized with richer Qwen descriptions produce stronger deletion and insertion effects than those optimized with

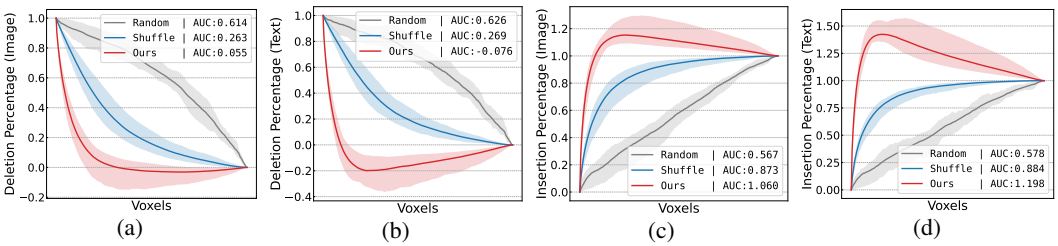


Figure 9: Deletion and insertion using a mismatched real fMRI volume as the replacement baseline. Panels (a,b) show deletion and (c,d) show insertion for image and text alignment, respectively. Random, Shuffle, and Ours use random rankings, the ranking from another sample's saliency map, and the ranking from the current sample's saliency map. The vertical axis reports alignment relative to the full input. Lower deletion AUC and higher insertion AUC indicate more influential rankings.

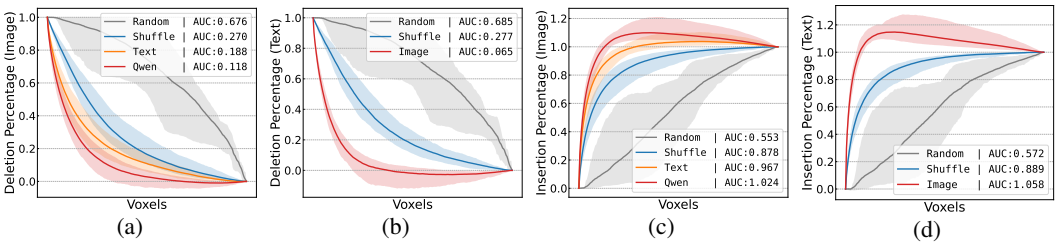


Figure 10: Cross-modal evaluation with different attribution and evaluation targets. Panels (a,c) evaluate image alignment using saliency maps optimized with COCO captions (Text) or Qwen descriptions (Qwen). Panels (b,d) evaluate text alignment using image-only attribution (Image). Panels (a,b) show deletion and (c,d) show insertion, with Random and Shuffle as controls. The vertical axis reports alignment relative to the full input.

COCO captions, consistent with more detailed descriptions retaining more of the paired image's semantic content. These results show that voxels selected using one modality also support alignment with the paired target in the other modality, extending the evidence beyond preservation of the exact attribution objective. Because both modalities occupy the same CLIP space, this remains a test of cross-modal consistency within the decoder, rather than independent evidence of neural causality.

E Computational Cost of Decision Attribution

We provide a rigorous analysis comparing the computational overhead of our gradient-based method against the standard perturbation-based baseline, RISE. Let $\mathcal{E}$ denote the Whole-Brain Decoder. We define the computational cost of a single forward pass as C_{fwd} and a backward pass (gradient computation) as C_{bwd}. Typically, $C_{\text{bwd}} \approx 2C_{\text{fwd}}$ in Transformer architectures (Kaplan et al., 2020).

• Baseline (RISE): RISE estimates saliency map by probing the model with random noise masks. To generate a high-fidelity saliency map, it requires a large number of masked forward passes (N) to ensure convergence. N is typically 1,000 or more, which is set to 1000 in our protocol. Crucially, RISE is inference-only and does not require gradient backpropagation. Thus, the total complexity is:

$$\text{Cost}_{\text{RISE}} = N \times C_{\text{fwd}} = 1000 \cdot C_{\text{fwd}}. \tag{14}$$

• Ours (IB-based Decision Attribution): Our method optimizes a parameterized saliency map via gradient descent. While each step requires both a forward and a backward pass, the strong semantic guidance from the Information Bottleneck objective allows for extremely rapid convergence. With only $N = 10$ optimization steps (a hyperparameter is tuned in Appendix F.4), the total complexity is:

$$\text{Cost}_{\text{Ours}} = N \times (C_{\text{fwd}} + C_{\text{bwd}}) \approx 10 \times (C_{\text{fwd}} + 2C_{\text{fwd}}) = 30 \cdot C_{\text{fwd}}. \tag{15}$$

F SUPPLEMENTARY RESULTS

F.1 ABLATION STUDY FOR WHOLE BRAIN DECODER

Table 2 presents ablation results for the whole-brain decoder on Subject 01.

Compared with the Visual ROIs decoder, the whole-brain decoder (Ours) improves all four retrieval directions, with clearer gains in image-to-brain and bidirectional text retrieval. These results show that whole-brain fMRI activity provides useful information for alignment with both visual and textual targets.

Replacing COCO captions with LLaVA or Qwen (Ours) descriptions improves all four retrieval directions. The improvement extends to image retrieval despite unchanged image targets, indicating that richer descriptions strengthen the shared brain-VLPM representation. The comparable image retrieval achieved by LLaVA and Qwen indicates that this gain reflects the empirical benefit of fine-grained descriptions and is not specific to a particular MLLM.

Table 2: Ablation study for the whole-brain decoder on Subject 01. Top-1 retrieval accuracy is reported as mean±std (%) over repeated retrieval trials. Visual ROIs evaluates a decoder restricted to predefined visual regions. COCO Caps and LLaVA-v1.6-13B evaluate alternative image descriptions. Ours denotes the whole-brain decoder with Qwen descriptions. **Best** results are highlighted.

Methods	Brain-Image	Image-Brain	Brain-Text	Text-Brain
Visual ROIs	70.30±2.03	54.50±2.63	53.71±2.79	43.27±2.25
COCO Caps	60.52±2.74	44.91±2.63	32.81±2.63	29.33±3.07
LLaVA-v1.6-13B	70.89±2.18	59.63±2.01	53.34±2.17	46.60±2.29
Ours	**71.58±2.08**	**59.99±2.14**	**57.53±2.36**	**49.48±2.14**

F.2 SUBJECT-WISE RESULTS

In this section, we provide a detailed breakdown of the experimental results for each individual subject. While the aggregate results demonstrate the overall effectiveness of our method, analyzing subject-specific data is crucial for assessing the stability and robustness of Brain-VLPM alignment across different individuals. Physiological differences and variations in signal quality can significantly impact fMRI decoding; therefore, consistent performance across all subjects serves as a strong validation of our framework. We report the granular metrics for both the retrieval tasks and the interpretability evaluations below.

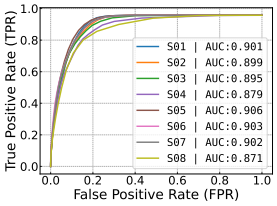


Figure 11: Macro-level ROI localization evaluated using ROC curves for all eight subjects. Curves closer to the upper-left corner indicate better discrimination between voxels inside and outside the nsd-general visual ROI.

1. Evaluation of Brain-VLPM Alignment. Tab. 3 reports the numerical results of cross-modal retrieval, providing a comparison against existing methods as well as the individual metrics for each of the 8 subjects.

2. Evaluation of Brain-VLPM Explanation. Fig. 15 shows the saliency maps generated by our method for the remaining 7 subjects. Fig. 11 and Fig. 12 present the subject-wise results for macro-level localization and micro-level ablation metrics, respectively. The visual-ROI agreement and deletion/insertion trends are consistent across subjects.

F.3 REPRESENTATIONAL SIMILARITY ANALYSIS

We employ Representational Similarity Analysis (RSA) to validate the semantic alignment. As shown in Fig. 13(a-c), the similarity patterns of generated brain embeddings closely mirror those of the ground-truth image and text embeddings. This qualitative alignment is confirmed by significant positive correlations in representational geometries (Fig. 13(d-e)), demonstrating that our framework effectively preserves the semantic relationships within the VLPM space.

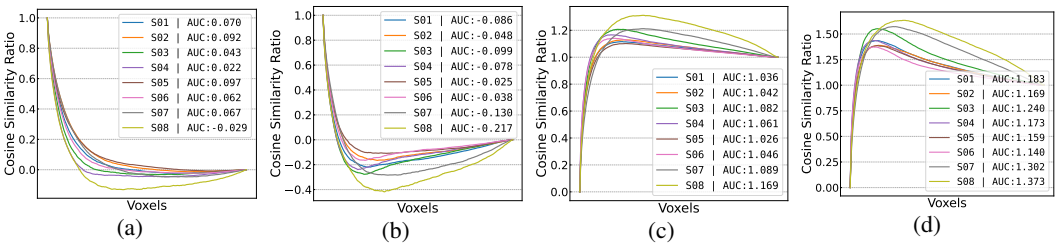


Figure 12: The micro-level ablation metric for each subject. The horizontal axis represents the voxels deleted (inserted) based on the saliency map, while the vertical axis represents the cosine similarity ratio of after ablation relative to before ablation baseline. (a-d) deletion and insertion for brain-image alignment (a,c) and brain-text alignment (b,d).

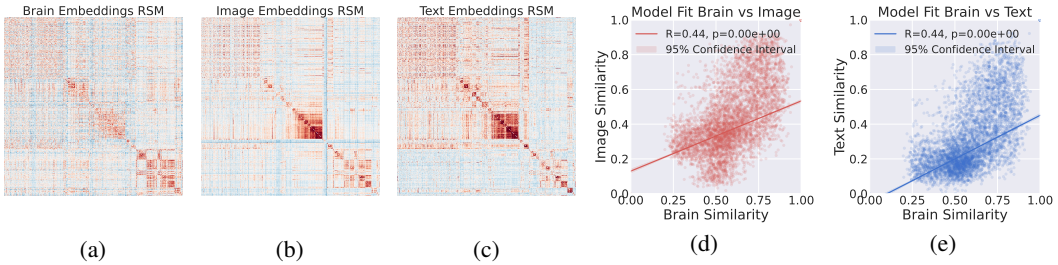


Figure 13: Visualization of Brain-VLPM alignment via Representational Similarity Analysis (RSA). (a-c) Representational Similarity Matrices (RSMs) visualizing the pairwise cosine similarities within Brain, Image, and Text embeddings. (d-e) Quantitative regression analysis comparing the Brain RSM against the Image RSM (d) and Text RSM (e). The dense clustering along the regression line indicates a high degree of structural correspondence between the decoder representations and the VLPM embeddings.

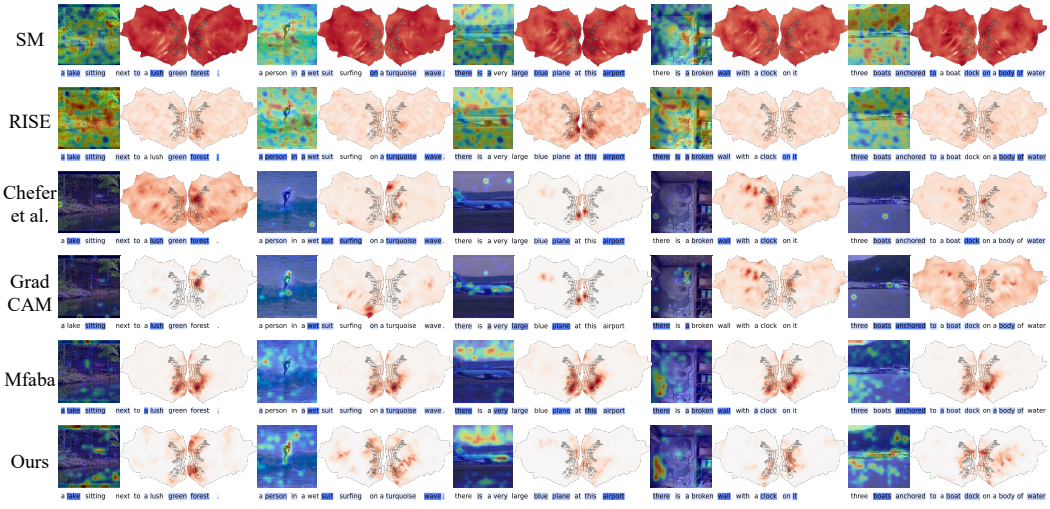


Figure 14: Compare the saliency maps of S01 for several explanation methods for brain-VLPM alignment. The heatmap shows voxel attribution scores for the corresponding alignment targets, highlighting voxels with high scores.

Table 3: Performance of brain multimodal retrieval accuracy (% mean ± std). Comparing ours with existing methods and detailing each of the 8 subjects. **Best** results are highlighted.

Methods	Brain-Image			Image-Brain			Brain-Text			Text-Brain		
	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10
Ridge	22.00 ±2.72	51.75 ±2.42	66.62 ±2.18	25.92 ±2.13	56.70 ±2.25	70.91 ±1.97	18.76 ±2.12	47.19 ±2.22	61.68 ±2.25	20.78 ±1.97	50.46 ±2.34	65.17 ±2.13
Mindeye	41.24 ±2.42	72.60 ±1.84	83.23 ±1.60	33.11 ±1.99	66.50 ±2.02	79.26 ±1.86	29.87 ±2.25	60.93 ±2.28	74.00 ±2.15	24.55 ±2.11	55.25 ±2.21	69.57 ±2.12
Mindeye2	36.86 ±2.03	67.85 ±2.07	79.29 ±1.70	32.89 ±2.14	65.35 ±2.18	78.02 ±1.92	27.08 ±2.17	58.18 ±2.33	71.34 ±2.12	24.35 ±2.24	55.14 ±2.42	69.22 ±2.15
TGBD	21.87 ±2.04	51.27 ±2.22	65.75 ±2.29	19.51 ±1.99	47.76 ±2.38	63.06 ±2.29	16.93 ±1.57	43.85 ±2.46	58.13 ±2.42	15.05 ±1.75	41.69 ±2.31	56.37 ±2.43
Ours	**57.40** ±2.22	**86.18** ±1.59	**93.29** ±1.18	**48.68** ±2.42	**81.30** ±1.79	**89.97** ±1.31	**44.60** ±2.31	**77.81** ±2.05	**88.01** ±1.50	**38.24** ±2.22	**73.11** ±2.07	**84.54** ±1.64
subj01	71.58 ±2.08	94.80 ±0.99	98.59 ±0.66	59.99 ±2.14	91.23 ±1.53	96.76 ±0.93	57.53 ±2.36	88.79 ±1.81	94.78 ±1.10	49.48 ±2.14	84.94 ±1.53	92.79 ±1.53
subj02	64.72 ±2.41	92.87 ±1.53	97.57 ±0.88	53.81 ±2.46	88.11 ±1.26	95.38 ±0.88	49.76 ±2.68	84.16 ±2.03	93.42 ±1.21	41.79 ±2.74	79.66 ±1.92	90.07 ±1.48
subj03	46.61 ±2.08	79.39 ±1.70	90.07 ±1.37	38.63 ±2.19	74.10 ±1.92	85.82 ±1.53	34.88 ±1.97	67.83 ±2.19	80.78 ±1.53	29.71 ±1.81	62.64 ±2.36	76.66 ±1.53
subj04	47.82 ±2.41	82.02 ±1.86	90.70 ±1.75	43.03 ±2.41	76.32 ±2.25	86.82 ±1.64	32.74 ±2.25	69.51 ±2.41	82.34 ±1.86	27.69 ±1.97	63.98 ±2.68	77.83 ±1.86
subj05	77.31 ±2.03	96.80 ±1.10	99.08 ±0.49	66.36 ±2.63	92.97 ±1.37	97.44 ±0.93	63.02 ±2.90	92.10 ±1.64	97.34 ±0.88	55.17 ±2.52	88.24 ±1.53	95.31 ±1.04
subj06	62.06 ±2.14	89.07 ±1.26	95.32 ±0.93	52.08 ±2.74	85.42 ±1.37	92.31 ±1.10	49.98 ±2.46	83.77 ±1.70	92.28 ±1.21	42.50 ±2.14	77.82 ±2.03	88.34 ±1.53
subj07	54.36 ±2.19	86.29 ±1.86	94.42 ±1.10	45.98 ±2.46	81.02 ±1.97	91.13 ±1.26	42.70 ±1.97	77.59 ±2.46	90.20 ±1.81	36.73 ±2.74	72.81 ±2.25	86.67 ±1.86
subj08	34.71 ±2.41	68.19 ±2.41	80.57 ±2.30	29.56 ±2.36	61.24 ±2.68	74.11 ±2.19	26.19 ±1.92	58.76 ±2.14	72.96 ±2.41	22.83 ±1.70	54.77 ±2.30	68.67 ±2.30

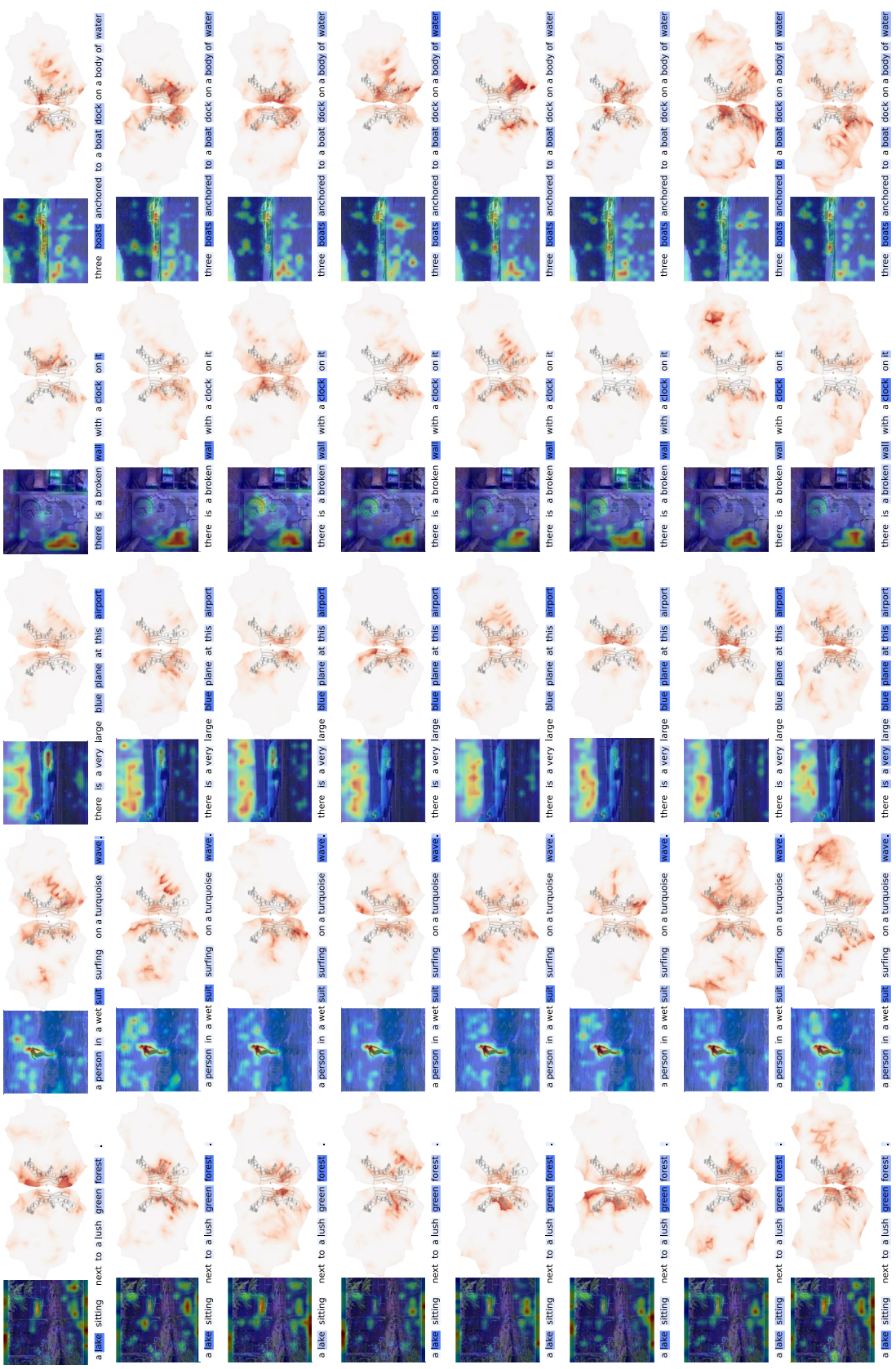


Figure 15: Saliency maps for other 8 subjects. The heatmap shows voxel attribution scores for the corresponding alignment targets, highlighting voxels with high scores.

F.4 Supplementary Hyperparameter Sensitivity Analysis

We conduct additional hyperparameter sensitivity analyses for our decision attribution method on participant 01, focusing on noise variance, learning rate, and optimization steps. We present examples of saliency maps generated with varying noise variances, learning rates, and optimization steps respectively in Fig. 16,17,18 and the quantitative results are presented in Fig. 19 for macro-level localization metrics, and Fig. 20,21,22 for micro-level ablation metrics.

Our analysis suggests that the saliency maps are stable across a wide range of hyperparameter settings. Specifically, varying the noise variance yields negligible differences in the resulting saliency maps, indicating stable performance across different noise levels. Regarding learning rate and optimization steps, we observe a slight trade-off: insufficient steps or low rates tend to produce denser, less discriminative maps, while excessive values may lead to over-sparsification. Nevertheless, ROI and deletion/insertion metrics vary little across these settings.

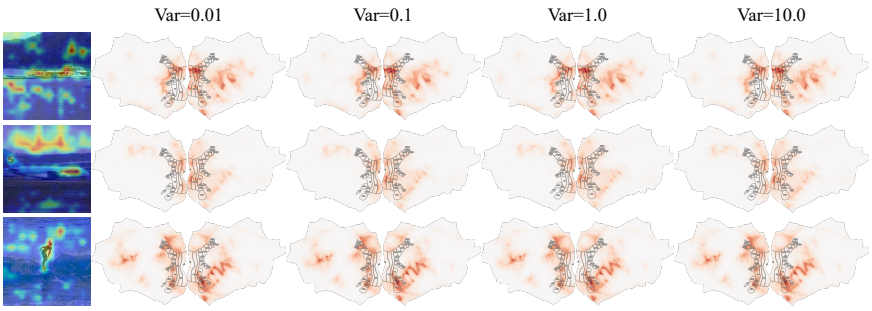


Figure 16: Examples of saliency maps of S01 obtained by different noise variances.

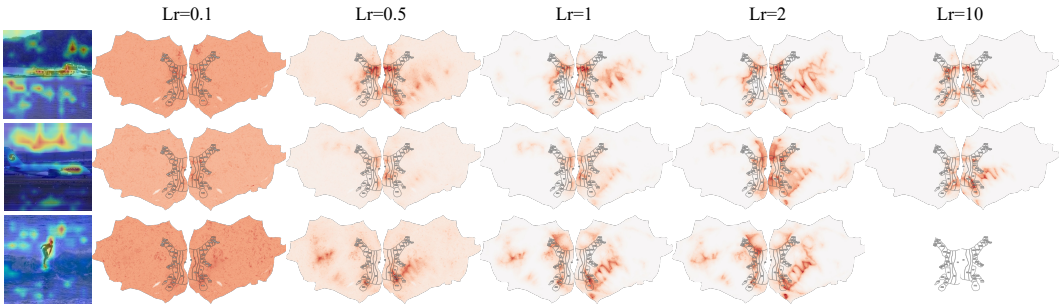


Figure 17: Examples of saliency maps of S01 obtained by different learning rates.

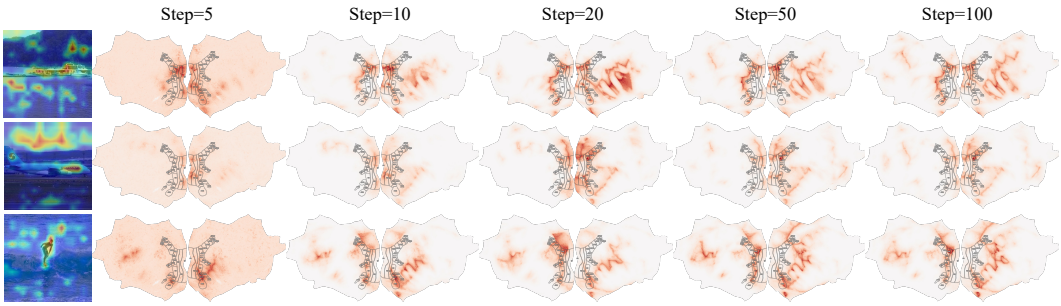
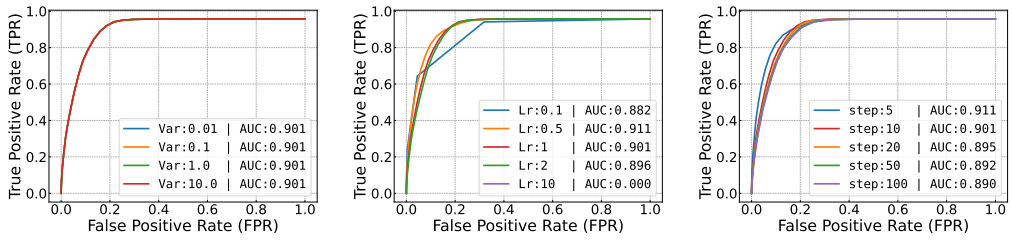
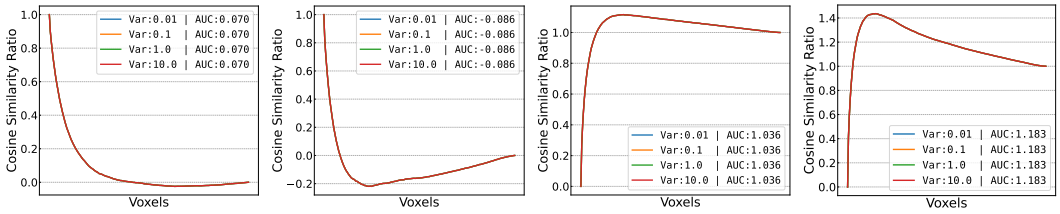


Figure 18: Examples of saliency maps of S01 obtained by different optimization steps.



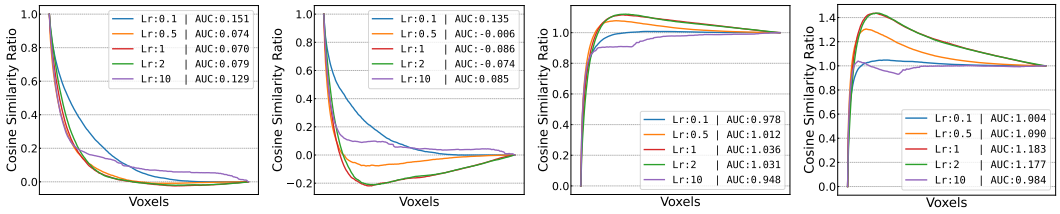
(a) Ablation of noise variance. (b) Ablation of learning rate. (c) Ablation of optimization steps.

Figure 19: The ROC-AUC of macro-level localization for different noise variances (a), learning rates (b), and optimization steps (c).



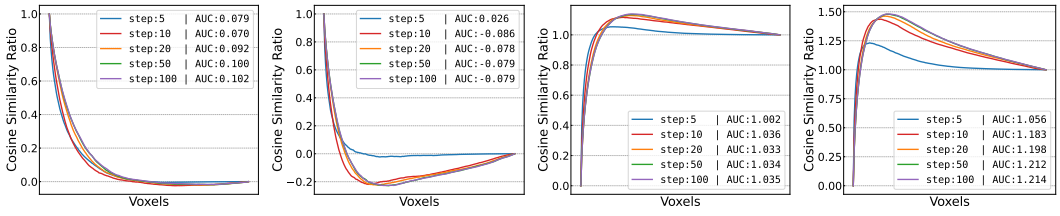
(a) (b) (c) (d)

Figure 20: The micro-level ablation metrics for different noise variances in our attribution method. Reporting the cosine similarity ratio of after/before deletion (insertion). (a-d) deletion/insertion voxels for brain-image alignment (a,c) and brain-text alignment (b,d).



(a) (b) (c) (d)

Figure 21: The micro-level ablation metrics for different learning rates in our attribution method. Reporting the cosine similarity ratio of after/before deletion (insertion). (a-d) deletion/insertion voxels for brain-image alignment (a,c) and brain-text alignment (b,d).



(a) (b) (c) (d)

Figure 22: The micro-level ablation metrics for different optimization steps in our attribution method. Reporting the cosine similarity ratio of after/before deletion (insertion). (a-d) deletion/insertion voxels for brain-image alignment (a,c) and brain-text alignment (b,d).

G Derivations and Proofs

G.1 Derivation of the Variational Information Bottleneck (Eq. 2)

The Information Bottleneck (IB) principle aims to maximize the mutual information between the latent representation Z and the target Y, while minimizing the mutual information between Z and the input X. The objective is given by:

$$\max I(Z;Y) - \beta I(Z;X). \tag{16}$$

Direct computation of mutual information is intractable in high-dimensional spaces. We approximate it using variational inference.

$$I(Z;Y) - \beta I(Z;X) \geq I_{\text{lower}}(Z;Y) - \beta I_{\text{upper}}(Z;X). \tag{17}$$

We maximize the right-hand side as a tractable variational surrogate for the original IB objective.

1. Upper Bound $I(Z;X)_{\text{upper}}$: We have $I(Z;X) = \int p(x,z) \log \frac{p(z|x)}{p(z)} dxdz$. Since the marginal $p(z) = \int p(z|x)p(x)dx$ is intractable, we introduce a fixed variational prior $r(z)$.

$$\begin{aligned} I(Z;X) &= \mathbb{E}_{p(x,z)}\left[\log \frac{p_\theta(z|x)}{p(z)}\right] \\ &= \mathbb{E}_{p(x)}\left[\mathbb{D}_{\text{KL}}[p_\theta(z|x)||r(z)]\right] - \mathbb{D}_{\text{KL}}[p(z)||r(z)] \\ &\leq \mathbb{E}_{p(x)}\left[\mathbb{D}_{\text{KL}}[p_\theta(z|x)||r(z)]\right]. \end{aligned} \tag{18}$$

Analytic KL Divergence. We assume the prior $r(z)$ is a standard spherical Gaussian $\mathcal{N}(0, I)$ and the posterior $p_\theta(z|x)$ is a multivariate Gaussian with a diagonal covariance structure $\mathcal{N}(\mu, \text{diag}(\sigma^2))$. Under these assumptions, the KL divergence term has a well-known analytical closed-form solution (Kingma & Welling, 2013):

$$\begin{aligned} \mathcal{L}_{KL} &= \mathbb{D}_{\text{KL}}[\mathcal{N}(\mu, \text{diag}(\sigma^2))||\mathcal{N}(0, I)] \\ &= -\frac{1}{2}\sum_{j=1}^{d}\left(1 + \log(\sigma_j^2) - \mu_j^2 - \sigma_j^2\right). \end{aligned} \tag{19}$$

2. Lower Bound $I(Z;Y)_{\text{lower}}$: By definition, $I(Z;Y) = \int p(z,y) \log \frac{p(y|z)}{p(y)} dzdy$. Since $p(y|z)$ is unknown, we introduce a variational approximation $q_\phi(y|z)$ to approximate the true posterior. Using the non-negativity of KL divergence $\mathbb{D}_{\text{KL}}[p(y|z)||q_\phi(y|z)] \geq 0$, we have:

$$\begin{aligned} I(Z;Y) &= \mathbb{E}_{p(z,y)}[\log p(y|z) - \log p(y)] \\ &= \mathbb{E}_{p(z,y)}[\log q_\phi(y|z)] + \mathbb{E}_{p(z)}[\mathbb{D}_{\text{KL}}[p(y|z)||q_\phi(y|z)]] + H(Y) \\ &\geq \mathbb{E}_{p(z,y)}[\log q_\phi(y|z)] + H(Y). \end{aligned} \tag{20}$$

Since $H(Y)$ is independent of the model parameters, maximizing this variational lower bound reduces to maximizing $\mathbb{E}_{p(z,y)}[\log q_\phi(y|z)]$.

G.2 Proof of the Fitting Term (Contrastive Alignment) (Eq. 3)

We aim to maximize the mutual information $I(Z;Y) = \mathbb{E}_{p(z,y)}[\log \frac{p(y|z)}{p(y)}]$. Direct computation is intractable due to the unknown densities. Following Oord et al. (2018), we estimate the density ratio $f(z,y) \propto \frac{p(y|z)}{p(y)}$ using a scoring function $f(z,y) = \exp(\text{sim}(z,y)/\tau)$. Given a batch of N samples containing one positive pair (z_i, y_i) and $N-1$ negative samples from the marginal $p(y)$, the optimal probability that the i-th sample is the positive one is:

$$p(d=i|Z,Y) = \frac{f(z_i,y_i)}{\sum_{j=1}^{N} f(z_i,y_j)} \approx \frac{\frac{p(y_i|z_i)}{p(y_i)}}{\sum_{j=1}^{N}\frac{p(y_j|z_i)}{p(y_j)}} = \frac{\exp(\text{sim}(z_i,y_i)/\tau)}{\sum_{j=1}^{N}\exp(\text{sim}(z_i,y_j)/\tau)}. \tag{21}$$

The corresponding one-way contrastive term is obtained by minimizing the negative log-likelihood of correctly identifying the positive sample:

$$\mathcal{L}_{Z\to Y} = -\mathbb{E}\left[\log \frac{f(z_i, y_i)}{\sum_{j=1}^{N} f(z_i, y_j)}\right] \approx -\mathbb{E}\left[\log \frac{\frac{p(y_i|z_i)}{p(y_i)}}{\sum_{j=1}^{N} \frac{p(y_j|z_i)}{p(y_j)}}\right]. \tag{22}$$

Using the inequality $\mathbb{E}[\log \frac{f}{\sum f}] \leq I(Z;Y) - \log(N)$ derived in Oord et al. (2018), we obtain the lower bound:

$$I(Z;Y) \geq \log(N) - \mathcal{L}_{Z\to Y}. \tag{23}$$

Thus, minimizing the one-way contrastive term $\mathcal{L}_{Z\to Y}$ maximizes a lower bound of the mutual information between two paired representations.

In our brain-VLPM alignment setting, the semantic representation Y is instantiated by image embeddings Y_{I} and text embeddings Y_{T}. For each semantic modality $M \in \{\mathrm{I}, \mathrm{T}\}$, Eq. 3 contains two one-way contrastive terms: the first term aligns brain decoder representations to modality-M embeddings, and the second term aligns modality-M embeddings back to brain decoder representations. Therefore, Eq. 3 averages the same contrastive lower-bound objective over two semantic modalities and two retrieval directions. Consequently, minimizing $\mathcal{L}_{\text{align}}$ jointly maximizes contrastive lower bounds of the mutual information between brain decoder representations and both image/text VLPM representations, serving as the fitting term that enforces semantic sufficiency in our Information Bottleneck objective.

G.3 Derivation of the Fitting Term of IB-based Decision Attribution (Eq. 9)

Here we derive the specific fitting term used in the IB-based Decision Attribution, justifying the use of cosine similarity as a proxy for Mutual Information maximization.

Following the VIB derivation above, we maximize the variational lower bound:

$$I(Z'_{\mathrm{B}}; Y_{\mathrm{I}}, Y_{\mathrm{T}}) \geq \mathbb{E}_{p(z'_{\mathrm{B}}, y_{\mathrm{I}}, y_{\mathrm{T}})}\left[\log q_\phi(y_{\mathrm{I}}, y_{\mathrm{T}}|z'_{\mathrm{B}})\right] + H(Y_{\mathrm{I}}, Y_{\mathrm{T}}). \tag{24}$$

Because the target entropy is independent of the attribution parameters, the expected log-likelihood serves as the fitting surrogate. We model the variational distribution as factorized across the image and text targets:

$$\log q_\phi(y_{\mathrm{I}}, y_{\mathrm{T}}|z'_{\mathrm{B}}) = \log q_\phi(y_{\mathrm{I}}|z'_{\mathrm{B}}) + \log q_\phi(y_{\mathrm{T}}|z'_{\mathrm{B}}). \tag{25}$$

We model the variational decoder $q_\phi(y|z)$ as an isotropic Gaussian distribution centered at the normalized brain decoder representation z'_{B}. Let y and z'_{B} be unit vectors (as is standard in CLIP embedding spaces, where features are ℓ_2-normalized). The probability density function is:

$$q_\phi(y|z'_{\mathrm{B}}) \propto \exp\left(-\frac{\|y - z'_{\mathrm{B}}\|^2}{2\sigma^2}\right), \tag{26}$$

where σ^2 is a variance parameter. Expanding the squared Euclidean distance for unit vectors $\|y\| = 1$ and $\|z'_{\mathrm{B}}\| = 1$:

$$\begin{aligned} \|y - z'_{\mathrm{B}}\|^2 &= (y - z'_{\mathrm{B}})^T(y - z'_{\mathrm{B}}) \\ &= \|y\|^2 + \|z'_{\mathrm{B}}\|^2 - 2y^T z'_{\mathrm{B}} \\ &= 1 + 1 - 2\cos(y, z'_{\mathrm{B}}) \\ &= 2 - 2\cos(y, z'_{\mathrm{B}}). \end{aligned} \tag{27}$$

Substituting this back into the log-likelihood:

$$\begin{aligned} \log q_\phi(y|z'_{\mathrm{B}}) &= -\frac{1}{2\sigma^2}(2 - 2\cos(y, z'_{\mathrm{B}})) + C_0 \\ &= \frac{1}{\sigma^2}\cos(y, z'_{\mathrm{B}}) + C. \end{aligned} \tag{28}$$

Here, σ^2 is fixed, and C_0 and C are independent of the attribution parameters. Thus, maximizing the log-likelihood is equivalent to maximizing the cosine similarity between the perturbed brain decoder representations and the target embeddings. This yields the fitting objective:

$$\max \mathbb{E}\left[\cos(y_{\mathrm{I}}, z'_{\mathrm{B}}) + \cos(y_{\mathrm{T}}, z'_{\mathrm{B}})\right]. \tag{29}$$

This derivation bridges the gap between the information-theoretic formulation and the cosine-similarity-based loss function employed in our methodology.

H GENERATE DETAILED TEXT DESCRIPTIONS

H.1 PROMPTING EXAMPLE FOR QWEN2.5-VL-32B

```
<|im_start|>system
You are an AI visual assistant that can analyze objects in the
    image. Currently, you receive an image and some sentences
    each describing the image you are observing. \n Please
    describe objects and relevance, concepts, background, color
    and scene of the image in a detailed manner but without
    decoration and embellishment. \n Always answer as if you are
    directly looking at the image. \n Describe the image content
    clearly and concisely and retain the meanings of each
    objects, relevance, concepts, background, color and scene in
    the image. \n Describe the image directly from the
    beginning. Do not with \'The image shows\' or \'The image
    depicts\'. Don\'t summarize or overall. \n Keep your answer
    less than 77 characters and words.
<|im_end|>

<|im_start|>user
{NSD captions of the image from the COCO dataset}
<|vision_start|><|image_pad|><|vision_end|>
<|im_end|>

<|im_start|>assistant
...
```

H.2 TEXT DESCRIPTIONS EXAMPLES FROM QWEN2.5-VL-32B



A person is kite surfing over ocean waves near the shore. They are airborne, performing a trick while holding onto the kite lines. The background includes houses and a clear sky. The water is choppy, indicating windy conditions.



A black and gray cat is sitting on the edge of a white toilet, facing left toward an open bathroom door. The scene includes a yellow mat on the floor and a shower curtain in the background.



A hand uses a knife to cut a homemade pizza on a wooden board. Next to it, there's a bowl of salad with utensils, two glasses of white wine, and two wine bottles on a wooden table. The scene suggests a casual dining setup.



A wooden table holds a white plate with a grilled sandwich (possibly a panini) and a side of green salad. Next to it is a blue Pepsi cup filled with dark soda. The background includes part of a person's arm and a tiled floor. The setting appears casual, likely a café or dining area.



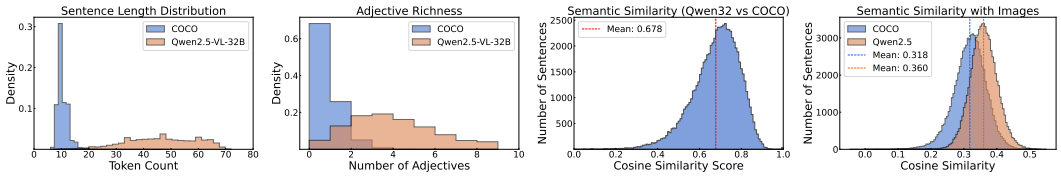
A man is crossing a street near an ice cream truck. The truck is white with colorful designs and text, parked on the right side of the road. The background includes urban buildings and a clear sky. The scene appears to be a busy intersection.

H.3 Retrieval Performance and Semantic Consistency of Generated Descriptions

We evaluate Qwen2.5-VL-32B descriptions through bidirectional CLIP retrieval and semantic consistency across descriptions of the same image.

1. Statistics of Generated Descriptions We conducted a statistical analysis of the Qwen2.5-VL-32B generated descriptions in comparison to the original COCO annotations. Fig. 23 presents 4 key metrics: (a) sentence length distribution, (b) adjective richness, (c) semantic similarity among descriptions for the same image, and (d) semantic similarity between descriptions and their corresponding images.

The results indicate that Qwen2.5-VL-32B generates descriptions that are more detailed and diverse. Specifically, the average sentence length is longer, and the adjective richness is higher compared to COCO annotations. Additionally, the semantic similarity analysis reveals that descriptions generated by Qwen2.5-VL-32B are more consistent with each other for the same image, while also showing higher similarity to the paired image embeddings in CLIP space. These findings suggest that the LLM-generated descriptions provide a richer and more nuanced representation of the visual stimuli, which is beneficial for downstream tasks such as brain-VLPM alignment.



(a) (b) (c) (d)

Figure 23: Statistical analysis of Qwen2.5-VL-32B generated descriptions against COCO annotations. (a) Distribution of sentence lengths in tokens. (b) Adjective richness measured as the average number of adjectives per description. (c) Semantic similarity distribution among descriptions for the same image by `all-MiniLM-L6-v2`. (d) Semantic similarity distribution between descriptions and their corresponding images by `CLIP-ViT-H-14`.

2. Bidirectional Cross-Modal Retrieval. We assessed discriminative specificity by projecting the 73,000 image-text pairs into a shared embedding space using CLIP-ViT-H/14. For each query, we calculate the cosine similarity between the query vector and all target vectors in the dataset. Retrieval performance is quantified using R@K, which measures the proportion of queries where the ground-truth target is ranked among the top K candidates based on similarity scores. Given the dataset scale of 73,000, the chance-level baseline is approximately 1.37×10^{-5} for R@1. As shown in Table 4, Qwen2.5-VL-32B substantially outperforms COCO annotations. In Text-to-Image (T2I) retrieval, which measures *discriminative specificity (how effectively a text description can distinguish its target image from 73,000 candidates)*, Qwen achieves a R@1 of 54.82% compared to COCO's 21.04%. Similarly, in Image-to-Text (I2T) retrieval, which evaluates whether image embeddings retrieve their paired descriptions, Qwen reaches 61.24% versus 30.89%. These results indicate that Qwen descriptions better distinguish the paired images and texts in CLIP space.

Table 4: Comparison of Bidirectional Retrieval Performance (N=73,000, chance level=K/73000). **Best** results are highlighted.

Direction	Annotation Source	R@1 (%)	R@5 (%)	R@10 (%)
Text-to-Image	Original COCO	21.04	38.64	47.10
	Qwen2.5-VL-32B	**54.82**	**76.31**	**82.45**
Image-to-Text	Original COCO	30.89	52.00	60.76
	Qwen2.5-VL-32B	**61.24**	**81.45**	**86.92**

3. Semantic Stability via Representation Convergence. We further measured stability by analyzing the retrieval success frequency across semantically diverse descriptions for the same stimulus. While COCO captions yield a mean retrieval success rate of 0.21 with 38,417 total failures (unidentifiable images), Qwen improves the average success rate to 0.54 and reduces the failure count to 25,961 (a

reduction of 12,456). These results indicate more consistent retrieval of the paired images across alternative descriptions in CLIP space.

In summary, Qwen2.5-VL-32B descriptions provide a discriminative and stable mapping to the original stimuli. The alignment metrics significantly outperform chance levels and demonstrate higher consistency in the CLIP feature space compared to original COCO annotations. These findings support using Qwen descriptions as discriminative text targets for brain-VLPM alignment.

I CONCEPT-CONDITIONED ATTRIBUTION

As illustrated in Fig. 24, we analyze a visual stimulus described as "a cat leans on the back of a bench." With the fMRI input and decoder fixed, we optimize separate saliency maps using "cat" and "bench" as the alignment targets. The "cat"-conditioned saliency map assigns high attribution scores to voxels around the Occipital Face Area (OFA) and lateral occipital regions (Kanwisher et al., 1997; Gauthier et al., 2000), whereas the "bench"-conditioned saliency map is more concentrated around the Parahippocampal Place Area (PPA) (Epstein & Kanwisher, 1998).

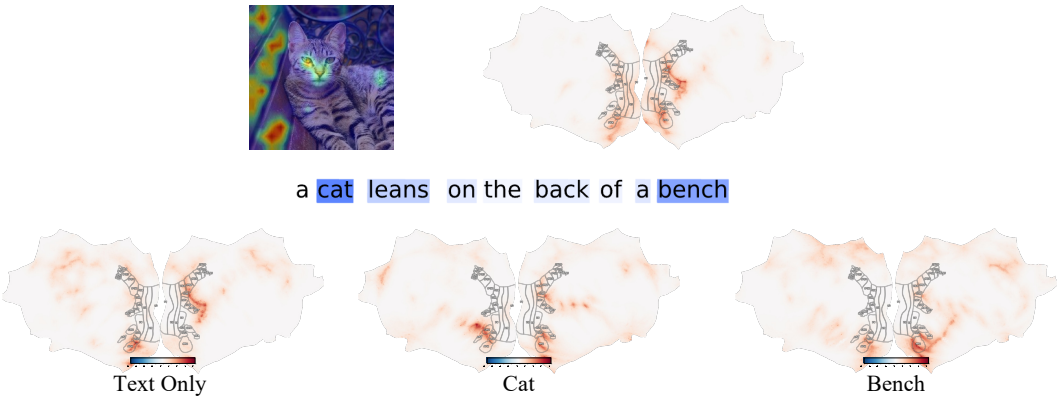


Figure 24: Concept-conditioned saliency maps for a fixed fMRI input and frozen decoder. The stimulus and image-conditioned saliency maps are shown above. The lower row compares maps optimized using the complete description, "cat," and "bench" as alignment targets. Only the alignment target changes across these maps.